Macular: a Multi-Task Adversarial Framework for Cross-Lingual Natural Language Understanding

Haoyu Wang
Purdue University
wang5346@purdue.edu

Yaqing Wang
Purdue University
wang5075@purdue.edu

Feijie Wu
Purdue University
wu1977@purdue.edu

Hongfei Xue
University at Buffalo
hongfeix@buffalo.edu

Jing Gao
Purdue University
jinggao@purdue.edu

ABSTRACT

Cross-lingual natural language understanding (NLU) aims to train NLU models on a source language and apply the models to NLU tasks in target languages, and is a fundamental task for many cross-language applications. Most of the existing cross-lingual NLU models assume the existence of parallel corpora so that words and sentences in source and target languages could be aligned. However, the construction of such parallel corpora is expensive and sometimes infeasible. Motivated by this challenge, recent works propose data augmentation or adversarial training methods to reduce the reliance on external parallel corpora. In this paper, we propose an orthogonal and novel perspective to tackle this challenging cross-lingual NLU task (i.e., when parallel corpora are unavailable). We propose to conduct multi-task learning across different tasks for mutual performance improvement on both source and target languages. The proposed multi-task learning framework is complementary to existing studies and could be integrated with existing methods to further improve their performance on challenging cross-lingual NLU tasks.

Towards this end, we propose a multi-task adversarial framework for cross-lingual NLU, namely Macular. The proposed Macular includes a multi-task module and a task-specific module to infer both the common knowledge across tasks and unique task characteristics. More specifically, in the multi-task module, we incorporate a task adversarial loss into training to ensure the derivation of task-shared knowledge only by the representations. In the task-specific fine-tuning module, we extract task-specific knowledge which is not captured by the multi-task module. A task-level consistency loss is added to the training loss so that consistent predictions across a target task and an auxiliary task (i.e., the task that is the most similar to the target task) are achieved. A language adversarial loss is also incorporated so that knowledge can be transferred from source languages to target ones. To validate the effectiveness of the proposed Macular, we conduct extensive experiments on four public datasets including paraphrase identification, natural language understanding, question answering matching, and query advertisement matching. The experimental results show that the proposed Macular can outperform state-of-the-art cross-lingual NLU approaches.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599864

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

cross-lingual, multi-task learning, natural language understanding

ACM Reference Format:
Haoyu Wang, Yaqing Wang, Feijie Wu, Hongfei Xue, and Jing Gao. 2023. Macular: a Multi-Task Adversarial Framework for Cross-Lingual Natural Language Understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3580305.3599864

1 INTRODUCTION

Natural language understanding (NLU), as an umbrella term, covers a variety of sub-tasks dealing with machine reading comprehension, including text classification, named entity recognition, and part-of-speech tagging. In this paper, we focus on NLU tasks in the cross-lingual setting, in which models are trained on source languages and then applied to target languages. Cross-lingual NLU is an important task for many cross-language applications. With the globalization trend, there is a great demand in the support of multiple languages by products and services. For example, Apple Siri and Amazon Alexa support 21 and 8 languages respectively, Google search engine supports 149 languages, and Instagram supports 36 languages. In these multi-language scenarios, there usually exist low-resource languages for which insufficient labeled samples are available. For a better understanding of those languages, it would be desirable to leverage the knowledge obtained from high-resource languages in a cross-lingual setting.

Cross-lingual NLU tasks are challenging. One major challenge is that labeled samples are usually unavailable for the target language (i.e., low-resource language). To tackle this challenge, methods have been proposed to align words or sentences of different languages in a low-dimensional embedding space [4, 15, 18, 32, 35] for knowledge transfer across languages. In recent years, the pre-trained multi-lingual large-scale language models, such as multilingual BERT [11] and XLM-RoBERTa [10], are developed and they achieve unattainable performance. Built on these multi-lingual language models, recent efforts focus on the inference of language-invariant representations [5, 7, 23, 24] and robust representations [14, 22].

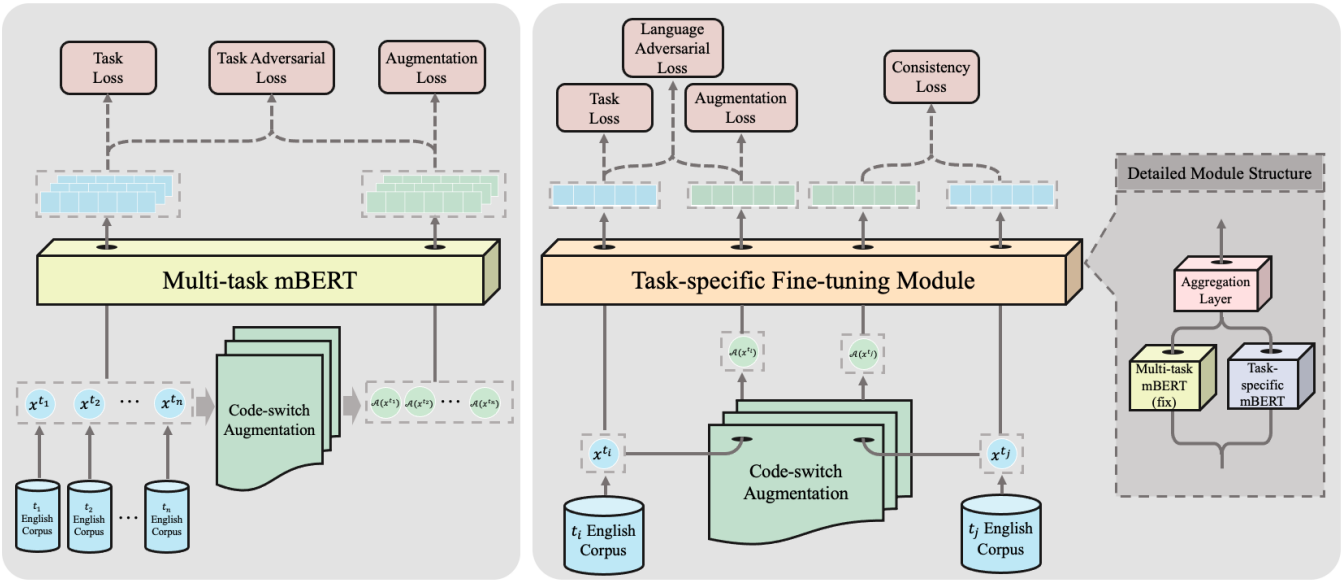

Figure 1: The framework of multi-task module (left) and task-specific fine-tuning module (right) of Macular.

Despite their outstanding performance, these existing approaches suffer from one major limitation, that is, they highly rely on parallel corpora, assuming that the translations of the same documents are available for multiple languages. Such corpora are very expensive to construct, especially for low-resource languages [22].

To overcome this limitation, augmentation methods have been proposed to reduce the reliance on parallel corpora. The augmentation is conducted according to the semantics of sentences or the syntax of languages [2]. For example, one popular augmentation approach is code switch augmentation [34], in which some words are randomly replaced in one sentence with its synonyms in other languages. In this paper, we propose a multi-task learning method that is complementary to existing cross-lingual NLU methods. We observe that different NLU tasks share some common properties, and thus multi-task learning enables the sharing of common knowledge across tasks and languages, and significantly reduces the requirement on large-scale labeled data. By the proposed multi-learning framework, the performance of NLU tasks on both source languages and target languages could be boosted.

Inspired by the fact that both unique and common task characteristics exist among different NLU tasks, the proposed framework is designed to learn task-shared representations and task-specific representations simultaneously. Then the intermediate representations of documents are a combination of task-shared and task-specific representations. By decoupling these two types of representations, we are able to capture both across-task shared knowledge and task-specific knowledge.

Based on this principle, we propose a multi-task adversarial framework for cross-lingual natural Language understanding, namely Macular, which includes a multi-task module and a task-specific fine-tuning module. The multi-task module is trained to learn the common representations via shared parameters of the language model across tasks. In the task-specific module, another language model is deployed to learn task-specific representations. Both representations are aggregated to make predictions. Different from existing works on cross-lingual NLU [5, 7, 23, 24], we do not require that parallel corpora are available. Instead, we work on corpora consisting of both high-resource and low-resource languages. Incorporating code switch augmentation [34] into the proposed multi-task learning framework can further reduce the reliance on parallel corpora. To capture the common characteristics across tasks and languages, we propose two corresponding adversarial losses that enable the multi-task representations to be invariant to tasks and enable the final representations to be invariant to languages. To further improve the performance, we propose a novel task-level consistency loss. For a target task, we pick another task which is the most similar, and then define consistency loss to enforce consistent predictions between the source language data and the augmented data of the chosen task. This consistency loss can help the model understand sentence semantics better, and can be considered as an alternative way of data augmentation that borrows information from a similar task.

The contributions of the paper are summarized as follows:

- We propose a novel multi-task adversarial framework for cross-lingual natural language understanding. To the best of our knowledge, this is the first attempt to leverage multi-task learning in the context of cross-lingual NLU tasks. We demonstrate that this effective strategy serves as a complementary way of data augmentation in cross-lingual NLU. With the proposed framework, parallel corpora are not needed and labeling efforts are greatly reduced.
- The proposed framework provides a new perspective of learning better cross-lingual representations by decoupling task-shared representations and task-specific representations. Novel models are proposed to obtain reasonable task-shared and task-specific

representations that capture the common and unique properties across tasks and languages respectively.
- We propose to integrate various losses into the training loss function to achieve the goal of cross-lingual NLU in a multi-task learning setting. In particular, a novel task-level consistency loss is proposed to help the model leverage information across tasks and languages.
- We conduct extensive experiments on four benchmark datasets. Results show that the proposed Macular outperforms baselines significantly. Furthermore, we study how the relationship among different NLU tasks affects the performance in ablation studies, which reveal insights and guidelines for multi-task learning in cross-lingual NLU.

2 RELATED WORK

2.1 Cross-lingual NLU

Cross-lingual NLU tasks are widely explored from two major perspectives: 1) word embedding [4, 15, 16, 18, 31, 32, 35, 37, 39, 40, 44], and 2) pre-trained multi-lingual language models, such as multilingual BERT [11], mT5 [45], MBart [30] and XLM-RoBERTa [10]. The state-of-the-art methods usually are based on pre-trained multilingual language models, i.e. fine-tuning language models on downstream task datasets. Recent works built upon multi-lingual language models can be briefly categorized as adversarial training-based and data augmentation-based models as discussed below.

Adversarial training-based cross-lingual models [1, 5–7, 19, 23, 24, 42, 47] are designed to learn language-agnostic representations. However, this line of work usually relies on the existence of parallel corpora or unlabelled target language data, which are expensive to obtain and not always available. To overcome the limitations, several recent adversarial training-based models [14, 22, 33] learn robust representations by enforcing consistent output over a neighborhood of data points from English which can be considered as adversarial examples [17] of English data. To better align the source language with target languages, some works explore data augmentation methods [2, 3, 12, 14, 34, 46] for cross-lingual NLU tasks. Various augmentation techniques were proposed. For example, the code-switch method [34] proposes to randomly replace the phrase in one sentence to other languages, and some other methods augment data via reordering [14], or based on syntax [2], or generate augmented data from the vicinity distribution of the source and target samples based on language models [3], or apply several existing data augmentation methods like code-switch [34] and machine translation jointly. Different from the aforementioned approaches, we propose a new perspective to tackle target cross-lingual NLU tasks via a multi-task learning framework. The proposed multi-task learning framework shows significant improvements on most of the languages in cross-lingual NLU tasks. It is also compatible with existing cross-lingual methods so that the performance can be further improved by an integration.

2.2 Multi-task Learning in Natural Language Understanding

Many multi-task learning models involve a shared feature extractor and task-specific output branches, such as [8, 9, 28, 29]. Other multi-task learning models [20, 36, 38] learn task sharing knowledge and task uniqueness by a hierarchical architecture, in which the lower layers focus on lower-level tasks like POS and higher layers focus on higher-level tasks like textual entailment. One existing work [26] proposes to introduce adversarial mechanism into multi-task learning such that learnt knowledge consists of both common patterns across tasks and unique task characteristics. However, this design is not suitable for large-scale language model since it needs to maintain a task-shared model as well as a task-specific module per task simultaneously during the training procedure, resulting in an extremely high memory requirement for training as the number of tasks increases. Different from this approach, the proposed Macular first builds one sharing multi-task module and then fine-tunes each task-specific module separately, reducing the training memory requirement largely. Moreover, all the aforementioned multi-task methods focus on NLU tasks in a monolingual setting, which is different from the cross-lingual setting studied in this paper.

3 METHODOLOGY

3.1 Problem Formulation

Cross-lingual natural language understanding (NLU) includes k tasks $t_1, t_2, ..., t_k$. For any task t_i, $i \in \{1, \ldots, k\}$, it only has the training data in the English language $\mathcal{D}_i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$, where x_j^i is the text content, y_j^i is the label of x_j^i, and n_i is the number of training instances in $\mathcal{D}_i$. The goal is to learn a model $f_i(x_j^i) \rightarrow y_j^i$ for each task t_i.

3.2 Overview

The objective is to learn an effective cross-lingual model which can be generalized from English to other target languages. Towards this goal, we propose a new **m**ulti-task **a**dversarial framework for **c**ross-ling**u**al natura**l** l**a**nguage unde**r**standing (Macular) to train a model consisting of two modules: (1) multi-task module, and (2) task-specific fine-tuning module. As shown in Fig. 1, Macular first learns task-shared representation via the multi-task module and then goes through the task-specific fine-tuning module for final predictions. The proposed framework does not require parallel corpora in different languages. Instead, it augments the original English-only dataset by code-switch augmentation, where the data include the texts of the source language and the augmented texts obtained by code-switch. Fig. 2 demonstrates code-switch augmentation with three vivid examples.

In the rest of the section, we briefly elaborate Fig. 1 and discuss how the training loss is defined for each module and how these two modules cooperate together. More details can be found in Section 3.3 and Section 3.4.

(1) **Multi-task Module (Section 3.3):** Its training loss consists of three parts, namely, task loss, augmentation loss, and task-adversarial loss. The first two losses aim to discover common knowledge across different tasks, and they are determined by the English corpus and the augmented data, respectively. Besides, we devise task-adversarial loss to learn task-invariant representation, which filters task-related information in multilingual BERT (mBERT) output.
(2) **Task-specific Fine-tuning Module (Section 3.4):** It comprises three components: a well-trained mBERT from the multi-task module to attain multi-task representations, a trainable

mBERT to learn task-specific representations, and an aggregation layer to aggregate both representations via a weighted sum. While training a task t_i, $i \in \{1, \ldots, k\}$, we fine-tune this module by minimizing the task loss, the augmentation loss, the language-adversarial loss, and the consistency loss. Specifically, the first two losses are computed with the English corpora and the augmented data of task t_i, respectively. Language-adversarial loss can sort out the unique information presented in the English corpus only and focus on the semantics of sentences. The consistency loss ensures that the predictions on task t_i are consistent with that on other similar tasks.

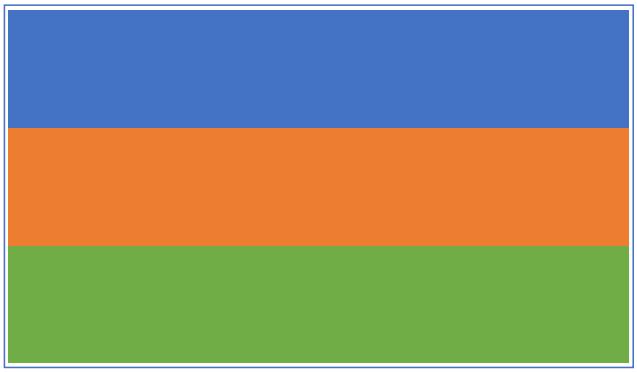

Figure 2: The example of code-switch augmentation. The blue font is en-de, orange font is en-fr, the purple font is en-zh, yellow font is en-bg, and the green font is en-ja.

3.3 Multi-task Module

In multi-task learning, it is uniqueness and commonness that coexist across different tasks. The multi-task module targets to capture common knowledge as multi-task representations. Therefore, it is essential to attain and purify the task-shared representation of a given text without mixing the task-specific representations.

Learn Multi-task Representations. To reduce the reliance on parallel corpora, we augment the training data of all the tasks with code-switch [34]. By this means, machine translation models are no longer needed, and the method is effective for augmentation in cross-lingual settings [34, 46]. In the multi-task setting, data augmentation can enlarge the training datasets, which benefits the model training because augmented data of different tasks can be shared. Fig. 2 exemplifies how code-switch works, and in each example, these two lines stand for the original and the augmented texts, respectively. Given j-th sample of task i ($i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, n_i\}$), let augmentation operation denote by $\mathcal{A}\left(x_j^i\right)$, and the augmented dataset be $\mathcal{D}_i^{aug} = \left\{\left(\mathcal{A}\left(x_j^i\right), y_j^i\right)\right\}_{j=1}^{n_i}$.

The multi-task module takes the pre-trained language model mBERT as the backbone. More detailedly, the module shares one mBERT encoder among different tasks and applies different prediction layers to different tasks. For a given training data point x_j^i, the mBERT $f_m(\cdot)$ encodes it to a d-dimensional vector:

$$\boldsymbol{e}_j^i = f_m\left(x_j^i\right). \tag{1}$$

The predictions made based on $\boldsymbol{e}_j^i$ are:

$$\hat{\boldsymbol{y}}_j^i = \sigma\left(\boldsymbol{W}^i \boldsymbol{e}_j^i\right) = F_m\left(x_j^i\right), \tag{2}$$

where we use σ to represent a sigmoid function while conducting binary classification or represent a softmax function while conducting multi-class classification. $\boldsymbol{W}^i$ is the prediction layer parameter for task t_i. Correspondingly, the task loss $\mathcal{L}_t$ and the augmentation loss $\mathcal{L}_a$ are formulated as

$$\mathcal{L}_t = \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} CE(F_m(x_j^i), y_j^i), \tag{3}$$

$$\mathcal{L}_a = \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} CE(F_m(\mathcal{A}(x_j^i)), y_j^i), \tag{4}$$

where $CE(\cdot, \cdot)$ represents cross-entropy loss function.

Purify Multi-task Representations. It is not a radical way that the module is trained by minimizing Eqn. 3 and 4 only because the unique knowledge of a task may mix up into task-shared representations. To ensure that the module can distinguish task-shared and task-specific representations, we propose a task-adversarial loss inspired by [26], which sorts out task-specific information via an adversarial game. More concretely, we set a discriminator $D_t(\cdot) : \mathbb{R}^d \to \{1, ..., k\}$ to predict the task IDs of the given training data points. If a representation only captures the common knowledge across tasks, then the discriminator hardly identifies a task ID. Of this motivation, we can optimize the discriminator by minimizing the following loss function:

$$\begin{aligned} &\frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} CE(D_t(f_m(x_j^i), \text{one\_hot}(i)) \\ &+ \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} CE(D_t(f_m(\mathcal{A}(x_j^i)), \text{one\_hot}(i)), \end{aligned} \tag{5}$$

where one_hot(i) represents a one-hot vector where i-th entry is 1 and the rest are 0.

Since the discriminator can predict the task IDs of the given data representations, we aim to deceive the discriminator with our mBERT encoder, where the discriminator fails in task identification. Specifically, we guide the encoder so that its output of a given input makes the discriminator have ambiguous predictions over task IDs. Formally, we aim to minimize the following task-adversarial loss

$$\begin{aligned} \mathcal{L}_{ta} =& \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} CE(D_t(f_m(x_j^i)), \frac{\mathbf{1}_k}{k}) \\ &+ \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} CE(D_t(f_m(\mathcal{A}(x_j^i)), \frac{\mathbf{1}_k}{k}). \end{aligned} \tag{6}$$

where $\mathbf{1}_k$ represents an all-one vector with k dimensions.

Training Process. The training of the multi-task module alternates with the training of the discriminator until the mBERT converges. As mentioned before, the discriminator minimizes Eqn. 5 to obtain the optimal parameters, while the multi-task module minimizes a compound loss

$$\mathcal{L}_m = \mathcal{L}_t + \mathcal{L}_a + \mathcal{L}_{ta}, \tag{7}$$

where $\mathcal{L}_t$, $\mathcal{L}_a$, and $\mathcal{L}_{ta}$ refer to the task loss (Eqn. 3), the augmentation loss (Eqn. 4), and the task-adversarial loss (Eqn. 6), respectively. While training the multi-task module, we freeze the discriminator to stabilize the procedure.

The multi-task training process is summarized in Alg. 1.

Algorithm 1: Multi-task Module

Input: Datasets of all tasks $\{\mathcal{D}_i\}_{i=1}^k$; training epoch t of the task discriminator.
Output: The learned multi-task module $F_m(\cdot)$.

```
// Do code-switch augmentation
for i ← 0 to k do
    D_i^aug = A(D_i)
Initialize mBERT F_m(·) and task discriminator D_t(·).
// Train F_m(·) and D_t(·)
while converge do
    // Update D_t(·)
    for i ← 0 to t do
        for batch in {{D_i}_{i=1}^k, {D_i^aug}_{i=1}^k} do
            Compute the task adversarial loss based on Eqn. 6;
            Update D_t(·);
    // Update F_m(·)
    for batch in {{D_i}_{i=1}^k, {D_i^aug}_{i=1}^k} do
        Compute the loss function for multi-task module according to Eqn. 7;
        Update F_m(·);
return F_m(·)
```

3.4 Task-specific Fine-tuning Module

The multi-task module is designed to capture commonness across tasks. In light of different languages having different grammatical structures, the multi-task representations are not sufficient to understand a cross-lingual input. Consequently, we design a task-specific fine-tuning module where one additional mBERT is introduced to extract the unique task representations. By freezing the well-trained multi-task module, we can train the newly introduced mBERT (denoted as $f_t(\cdot)$) and realize the trade-off between the task-shared and the task-specific representations. The following subsections discuss how to achieve these two goals.

3.4.1 Aggregation Layer. The aggregation layer is to aggregate the task-shared representation from the frozen multi-task module and the task-specific representation from the task-specific mBERT. Inspired by [41], we aggregate these two representations in a weighted-sum way. Formally, for a given data point x_j^i, the aggregated representation is written as

$$\boldsymbol{v}_j^i = \text{Agg}(f_m(x_j^i), f_t(x_j^i)) = \alpha f_m(x_j^i) + (1-\alpha) f_t(x_j^i), \tag{8}$$

where α is a hyper-parameter to balance the weights between the multi-task representation $f_m(x_j^i)$ and the task-specific representation $f_t(x_j^i)$. The final prediction is output via a fully-connected layer:

$$\hat{\boldsymbol{y}}_j^i = \sigma(\boldsymbol{M}^i \boldsymbol{v}_j^i) = F_t(x_j^i), \tag{9}$$

where σ represents a sigmoid function while conducting binary classification or represents a softmax function for multi-class classification. Let $\boldsymbol{M}^i$ denote the prediction layer parameter for task t_i. As the multi-task module $f_m(\cdot)$ is frozen, the training solely updates the task-specific mBERT $f_t(\cdot)$.

3.4.2 Loss Function. The training of the task-specific fine-tuning module relies on four losses: (1) task loss, (2) augmentation loss, (3) language adversarial loss, and (4) consistency loss. The first two losses are similar to those defined in the multi-task module, which can be presented as

$$\mathcal{L}_t = \frac{1}{n_i} \sum_{j=1}^{n_i} CE(F_t(x_j^i), y_j^i), \tag{10}$$

$$\mathcal{L}_a = \frac{1}{n_i} \sum_{j=1}^{n_i} KL_s(F_t(\mathcal{A}(x_j^i), F_t(x_j^i)), \tag{11}$$

where $KL_s(p,q) = KL(\text{stop\_grad}(p), q) + KL(p, \text{stop\_grad}(q))$ is the symmetrical Kullback-Leibler divergence [46]. The $KL_s(\cdot)$ encourages consistent predictions between the English texts and their own augmented texts.

The language-adversarial loss is designed to encourage the model to learn transferable representation across languages instead of focusing on language characteristics. Existing studies [5, 7, 23, 24] have explored adversarial loss in the cross-lingual setting when parallel or translation corpora are available. However, considering the unavailability of parallel corpora, we cannot adopt their solutions. In replacement, we utilize our augmented data with non-English words and follow the motivation of adversarial loss to generate language-invariant representations. To be more specific, we define a language discriminator $D_l(\cdot) : \mathbb{R}^d \rightarrow \{0, 1\}$ to distinguish English text and augmented text, which minimizes the following language classification loss

$$\frac{1}{n_i} \sum_{j=1}^{n_i} CE(D_l(\text{Agg}(f_m(x_j^i), f_t(x_j^i))), \text{one\_hot}(0)) + \frac{1}{n_i} \sum_{j=1}^{n_i} CE(D_t(\text{Agg}(f_m(\mathcal{A}(x_j^i)), f_t(\mathcal{A}(x_j^i))), \text{one\_hot}(1)). \tag{12}$$

The task-specific mBERT $f_t(\cdot)$ is to learn language-invariant representations by fooling $D_l(\cdot)$. In other words, we encourage $f_t(\cdot)$ to prevent the language discriminator from identifying whether or not the text is in English. As a result, we formulate the following

language-adversarial loss

$$\mathcal{L}_{la} = \frac{1}{n_i}\sum_{j=1}^{n_i} CE(D_l(\text{Agg}(f_m(x_j^i), f_t(x_j^i)), \frac{1}{2}) + \frac{1}{n_i}\sum_{j=1}^{n_i} CE(D_l(\text{Agg}(f_m(\mathcal{A}(x_j^i)), f_t(\mathcal{A}(x_j^i))), \frac{1}{2}). \quad (13)$$

In the end, we introduce a novel task-level consistency loss, which is to encourage consistent predictions on a similar task to the target task of interest. This loss is helpful in learning smooth representation and capturing generalized patterns. In detail, for a given task t_i, we first pick a task t_j which is similar to t_i. For example, if task t_i is PI, we can pick NLI task as task t_j. Then we encourage the model predictions on the augmented text of t_j and English text of t_j to be consistent. Formally, we minimize the following consistency loss

$$\mathcal{L}_c = \frac{1}{n_j}\sum_{k=1}^{n_i} KL_s(F_t(\mathcal{A}(x_k^j), F_t(x_k^j)). \quad (14)$$

The above four loss functions are combined to form the final loss function of the task-specific fine-tuning module as follows:

$$\mathcal{L}_s = \mathcal{L}_t + \beta_1\mathcal{L}_a + \beta_2\mathcal{L}_{ta} + \beta_3\mathcal{L}_c, \quad (15)$$

where β_1, β_2, and β_3 are hyper-parameters. We summarize the entire training process of task-specific fine-tuning in the Alg. 2.

Algorithm 2: Task-specific Fine-tuning Module

Input: Datasets of current task and the chosen task $\mathcal{D}_i$, $\mathcal{D}_j$, $\mathcal{D}_i^{aug}$, $\mathcal{D}_j^{aug}$; training epoch t of the language discriminator; coefficients of different loss β_1, β_2, and β_3; trained multi-task module $F_m(\cdot)$.

Output: The learned task-specific fine-tuning module $F_t(\cdot)$.

Initialize $F_t(\cdot)$ and language discriminator $D_l(\cdot)$.
// Train $F_t(\cdot)$ and $D_l(\cdot)$
while *converge* **do**
  // Update $D_t(\cdot)$
  for $i \leftarrow 0$ **to** t **do**
    for *batch in* $\{\mathcal{D}_i, \mathcal{D}_i^{aug}, \mathcal{D}_j, \mathcal{D}_j^{aug}\}$ **do**
      Compute the language adversarial loss based on Eqn. 12;
      Update $D_l(\cdot)$;
  // Update $F_t(\cdot)$
  for *batch in* $\{\mathcal{D}_i, \mathcal{D}_i^{aug}, \mathcal{D}_j, \mathcal{D}_j^{aug}\}$ **do**
    Compute the loss function for ask-specific fine-tuning module according to Eqn. 15;
    Update $F_t(\cdot)$;
return $F_t(\cdot)$

4 EXPERIMENT

In this section, we evaluate the proposed Macular with the goal of answering the following questions.

- RQ1 How does Macular perform compared to state-of-the-art baselines?
- RQ2 What are the roles of task adversarial loss, language adversarial loss and consistency loss in model performance improvements respectively?
- RQ3 How does the performance change with respect to different chosen tasks for multi-task learning?
- RQ4 Can the proposed Macular be generalized to other backbones?

4.1 Datasets and Experiment Settings

4.1.1 Datasets. To fairly evaluate the performance of the proposed model, we conduct experiments on four public benchmark datasets including PAWS-X [21], XNLI [21], QAM [25], and QADSM [25]. These four datasets corresponds to four tasks as paraphrase identification (PI), natural language inference (NLI), question answering matching (QA matching), and query advertisement matching (QAD matching) respectively. Each dataset is in several languages. For training and validation set, we only have data in English following the existing work [34]. The developed models are then evaluated on test dataset in multiple languages. The statistics of datasets is summarized in Table 1.

Table 1: Statistics of datasets.

Dataset	# of languages	Task	$\|\text{Train}\|^{en}$	$\|\text{Dev}\|^{en}$	$\|\text{Test}\|^{avg}$
PAWS-X	7	PI	49k	2k	14k
XNLI	15	NLI	392k	2k	5k
QAM	3	QA matching	100k	10k	10k
QADSM	3	QAD matching	100k	10k	10k

4.1.2 Baselines. We adopt six state-of-the-art methods as baselines:

- mBERT [11] is a multi-lingual version of BERT and is pre-trained on Wikipedia corpora in 104 most widely used languages. mBERT is one of the state-of-the-art methods for cross-lingual natural language processing tasks.
- Prompt [13] is one of the state-of-the-art fine-tuning paradigms ans shows great performance in various NLP tasks by reformulating the classification tasks into a fill-in-the-blank format. Recent work [27] shows prompt-based methods have the strong transferable ability on English text. We follow their setting and introduce this into the cross-lingual setting.
- RS-DA [22] is a state-of-the-art cross-lingual approach based on robust learning. It forces the model to make similar predictions for representation in the same robust region. We adopt their official release of RS-DA implementation [1].
- Syn. [2] is a state-of-the-art cross-lingual approach which encodes the universal dependency tree structure in mBERT to conduct cross-lingual transfer. We adopt the official code release of Syn.[2].
- CoSDA-ML [34] is a state-of-the-art cross-lingual approach which introduces code-switch augmentation into cross-lingual tasks.

To study the role of the task-specific fine-tuning proposed in subsection 3.4, we propose a reduced model Macular-NoTS, which

[1] https://github.com/uclanlp/Robust-XLT
[2] https://github.com/wasiahmad/Syntax-MBERT

Table 2: Performance comparison on the four datasets. "AVG" means the average accuracy of all languages. The highest scores per category are in bold. Results of ⋆ are taken from [25], and results of * are taken from [21]. Results of ♡ are taken from [22] and [2] or obtain based on their official code release.

Task	Method	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	ja	ko	AVG
QADSM	mBERT*	68.3	-	-	60.3	-	-	64.1	-	-	-	-	-	-	-	-	-	-	64.2
	Prompt	67.2	-	-	59.4	-	-	62.5	-	-	-	-	-	-	-	-	-	-	63.1
	RS-DA [22]♡	67.4	-	-	58.5	-	-	61.0	-	-	-	-	-	-	-	-	-	-	62.3
	Syn. [2]♡	68.4	-	-	60.8	-	-	64.0	-	-	-	-	-	-	-	-	-	-	64.4
	CoSDA-ML	68.2	-	-	60.1	-	-	63.4	-	-	-	-	-	-	-	-	-	-	63.9
	Macular-NoTS	69.8	-	-	61.0	-	-	65.5	-	-	-	-	-	-	-	-	-	-	65.4
	Macular	**69.9**	-	-	**62.0**	-	-	**66.2**	-	-	-	-	-	-	-	-	-	-	**66.0**
QAM	mBERT*	67.5	-	-	64.7	-	-	66.0	-	-	-	-	-	-	-	-	-	-	66.1
	Prompt	68.7	-	-	64.0	-	-	63.8	-	-	-	-	-	-	-	-	-	-	65.4
	RS-DA [22]♡	66.5	-	-	61.5	-	-	62.9	-	-	-	-	-	-	-	-	-	-	63.6
	Syn. [2]♡	68.8	-	-	63.6	-	-	64.7	-	-	-	-	-	-	-	-	-	-	65.7
	CoSDA-ML	69.5	-	-	64.1	-	-	65.5	-	-	-	-	-	-	-	-	-	-	66.4
	Macular-NoTS	69.2	-	-	64.4	-	-	**66.5**	-	-	-	-	-	-	-	-	-	-	66.7
	Macular	**69.7**	-	-	**64.9**	-	-	66.1	-	-	-	-	-	-	-	-	-	-	**66.9**
PAWS-X	mBERT⋆	94.0	-	-	85.7	-	87.4	87.0	-	-	-	-	-	-	-	77.0	73.0	69.6	82.0
	Prompt	94.0	-	-	85.8	-	88.6	87.6	-	-	-	-	-	-	-	79.5	75.4	74.8	83.7
	RS-DA [22]♡	93.5	-	-	87.8	-	88.8	88.8	-	-	-	-	-	-	-	81.5	79.3	78.3	85.4
	Syn. [2]♡	94.0	-	-	87.8	-	85.9	89.1	-	-	-	-	-	-	-	80.7	75.8	76.3	84.3
	CoSDA-ML	94.4	-	-	87.0	-	89.8	**89.3**	-	-	-	-	-	-	-	82.7	78.7	**79.7**	85.9
	Macular-NoTS	95.1	-	-	**88.2**	-	89.4	88.8	-	-	-	-	-	-	-	83.5	79.9	78.8	86.2
	Macular	**95.2**	-	-	88.1	-	**90.0**	**89.3**	-	-	-	-	-	-	-	**83.6**	**80.3**	79.0	**86.5**
XNLI	mBERT⋆	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	-	-	65.4
	Prompt	81.3	63.6	67.9	69.6	67.1	73.3	72.0	59.7	67.1	51.1	54.2	61.4	58.1	69.5	68.2	-	-	65.6
	RS-DA [22]♡	81.0	66.4	69.9	71.8	68.0	74.7	74.2	62.7	70.6	51.1	55.7	62.9	60.9	71.8	71.4	-	-	67.6
	Syn. [2]♡	81.6	65.4	69.3	70.7	66.5	74.1	73.2	60.5	68.8	-	-	62.4	58.7	69.9	69.3	-	-	68.5
	CoSDA-ML	**82.8**	**68.2**	71.9	72.5	70.0	76.7	75.4	64.9	**72.3**	50.5	58.6	**63.9**	60.7	**73.2**	72.8	-	-	68.9
	Macular-NoTS	81.5	68.1	71.2	72.8	69.9	75.7	74.0	64.2	71.2	**51.5**	59.0	63.6	61.2	71.8	72.6	-	-	68.5
	Macular	82.6	67.9	**72.2**	**73.8**	**70.5**	**76.8**	**75.5**	**65.0**	**72.3**	51.3	**59.5**	63.8	**61.8**	72.4	**73.0**	-	-	**69.2**

only use multi-task module introduced in Section 3.3 without the task-specific fine-tuning module.

4.1.3 Evaluation Metric and Implementation Details. Following existing works[21, 25], we evaluate four tasks including paraphrase identification, natural language inference, question answering matching, and query advertisement matching in term of Accuracy. We implement the backbone based on the Huggingface Transformers library[3] [43]. We tune the batch size and learning rate on the validation set via a grid search over $\{4, 8, 16, 32\}$ and $\{1e-6, 5e-6, 1e-5, 2e-5, 3e-5, 5e-5\}$ respectively. For the coefficient α in Eqn. 8, we tune it on the validation set via a grid search over $\{0.1, 0.2, ..., 0.9\}$. We set $\beta_1 = 1$, $\beta_2 = \beta_3 = 0.1$ for all tasks. The training epochs of discriminators are set as 2. We use a one hidden layer deep neural network using ReLU activation as the discriminator. All experiments are repeatedly run 3 times and the corresponding average results are reported. We run our experiments on the server with 4 NVIDIA RTX A6000 and an Intel Xeon Gold 6254 CPU.

4.2 Performance Comparison

In this section, we report the performance of baselines and the proposed Macular in Table 2 to answer RQ1.

Table 2 shows that the proposed Macular outperforms all the state-of-the-art baselines on four tasks in terms of average accuracy over several languages. The best baselines on four datasets are Syn. and CoSDA-ML, which target cross-lingual tasks by data augmentation from the perspectives of syntax and semantics. Such an observation further confirms the effectiveness of data augmentation in cross-lingual tasks. The proposed Macular improves cross-lingual task performance by incorporating multi-task mechanism. The general improvements are observed on different languages and different tasks. Take QADSM task as an example, the proposed Macular brings improvements around 2%, 2.0%, 3.4% compared to the best baselines over English, German and French languages. The effectiveness of multi-task learning mechanism can be also justified by the performance of Macular-NoTS. As a reduced version of the proposed framework Macular, Macular-NoTS reduces task-specific module and only keeps multi-task module. Such a model is able to outperform the state-of-the-art baselines on most of tasks and achieve comparable performance on XNLI. After adding back task-specific fine-tuning module, Macular is able to integrate task-shared and task-specific representations to achieve better performance compared to Macular-NoTS on all the tasks in term of average accuracy over languages.

[3]https://github.com/huggingface/transformers

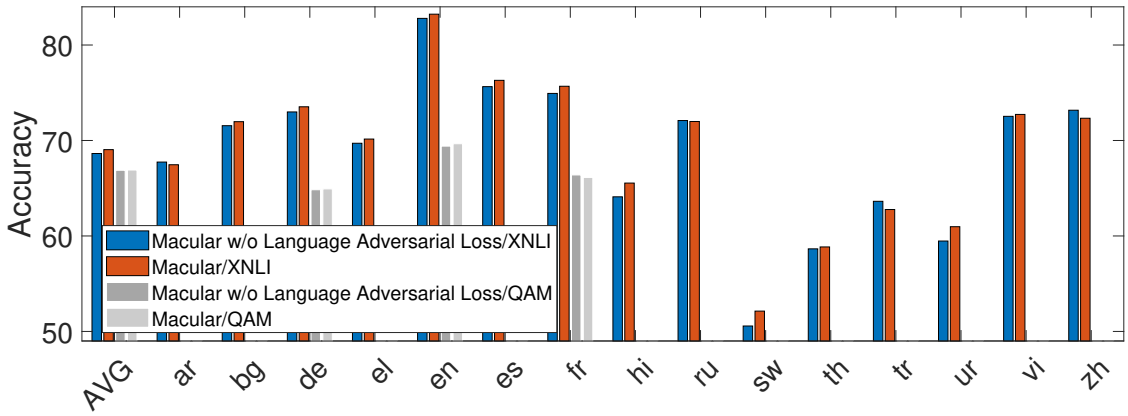

Figure 3: Accuracy of Macular with or without the proposed language adversarial loss on XNLI and QAM dataset. The blue and orange bars show the results on the XNLI dataset, and gray bars show the results on the QAM dataset.

4.3 Effectiveness of Task Adversarial Loss

In this section, we do the ablation study to answer RQ2 regarding task adversarial loss. To intuitively illustrate the role of task adversarial loss in the proposed model, we design two ablation studies by removing task adversarial loss from the reduced model Macular-NoTS and the proposed framework Macular. The performance comparison between Macular-NoTS without task adversarial loss and Macular-NoTS is able to help us tell if the task adversarial loss can help the multi-task module to learn better task-shared representation. We further study the effect of the task adversarial loss on the task-specific fine-tuning module via comparing Macular without task adversarial loss and Macular. The corresponding experiments are conducted on XNLI and QAM datsets corresponding to inference and QA matching tasks as examples. We show these results in Fig. 4.

First, the task adversarial loss is shown to be effective for the multi-task module to learn better task-shared representation based on performance comparison between Macular-NoTS and Macular-NoTS w/o task adversarial loss. The Macular-NoTS outperforms Macular-NoTS w/o task adversarial loss on QAM and XNLI. One possible explanation behind this is that task adversarial loss can help model to learn transferable patterns shared across different tasks and transferable patterns can further generalise to different languages. Second, the task adversarial loss is also confirmed to help boost the performance of task-specific fine-tuning module. Fig. 4 shows Macular outperforms Macular without task adversarial loss. For Macular without task adversarial loss, the representation learned from multi-task module also contains task-specific information, which may not learn transferable representation, thereby degrading performance.

4.4 Effectiveness of Language Adversarial Loss

In this section, we do the ablation study to answer the RQ2 regarding langauge adversrial loss. Similar to Section 4.3, we design an ablation study by removing language adversarial loss from the proposed framework Macular to intuitively illustrate the role of language adversarial loss in the proposed model. The performance comparison between Macular and Macular without the language adversarial loss on XNLI and QAM datasets are reported in Fig. 3.

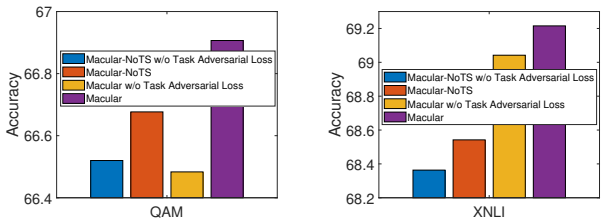

Figure 4: Accuracy of Macular-NoTS and Macular with or without the proposed task adversarial loss on XNLI and QAM dataset.

Fig. 3 shows that the average accuracy of Macular is higher than that of Macular without language adversarial loss on two datasets, confirming the positive role of language adversarial loss in the proposed framework Macular. Language adversarial loss is designed to help the model learn more transferable representation instead of language-specific patterns. However, due to lacking parallel corpora in different languages, the discriminator faces more difficulties in learning different language patterns and thus the improvement from the language adversarial loss in the proposed framework Macular is less significant compared to improvements reported in the existing work [24].

Table 3: Performance comparison of Macular and Macular w/o consistency loss.

	QADSM	QAM	PAWS-X	NLI
Macular w/o Consistency Loss	65.6	66.7	86.4	68.3
Macular	66.0	66.9	86.5	69.2

4.5 Effectiveness of Consistency Loss

In this section, we do the ablation study to answer RQ2 regarding consistency loss. We show the performance comparison between Macular and Macular without consistency loss to study the effect of the consistency loss in the proposed model. The results are reported

in Table 3. Table 3 shows that Macular outperforms Macular without consistency loss on the four datasets. The proposed consistency loss is able to leverage other similar task data to further augment the data of target task and enforce a smooth representation for better transferable ability.

Table 4: Performance comparison of multi-task module with different task combinations.

	QADSM	QAM	PAWS-X	XNLI
Macular-NoTS w/o QADSM	-	66.0	86.2	68.4
Macular-NoTS w/o QAM	64.7	-	86.3	68.7
Macular-NoTS w/o PAWS-X	65.4	66.2	-	68.5
Macular-NoTS w/o XNLI	64.9	66.2	85.9	-
Macular-NoTS	65.4	66.7	86.2	68.5

4.6 Sensitivity w.r.t. Chosen Tasks for Multi-task Training

In this section, we study the effect of task choices on multi-task module performance to answer RQ3. To show the differences of tasks, we conduct experiments in four task settings, where we remove one task from the four tasks and the removed task is different in each time. We show the performance comparison of four settings to analyze the effect of each task in Table 4. First, multi-task training with four tasks achieves best performance compared to newly proposed four settings. This observation confirms the importance of each task. In a more fine-grained level, NLI task is able to make general contribution to the performance of other tasks. Compared to full version of multi-task, multi-task learning without NLI leads to general performance drops since NLI task is to infer logic relationship among sentences and is a fundamental task to others. We also observe that QADSM and QAM can benefit each other significantly. This is because QAM and QADSM both need to predict if the last sentence matches the previous one and share lots of similarities. However, we also find QADSM and QAM can not provide a large performance boost for PI and NLI, and QAM may degrade the performance of PI and NLI. The reasons may lie in two folds: first one is that the QAM and QADSM tasks are to predict matching relationship which is different from inference task in PI and NLI. Another one is that the data distributions of them are different. QAM and QADSM usually have short questions but long answers while the sentence length in PI and NLI is usually similar.

4.7 Extension to Other Backbone

In this section, we study the generalization ability to answer RQ4. Similar to Section 4.3, we show the performance of Macular with XLM-RoBERTa-base (XLM-R) [10] as the backbone on the XNLI and QAM datasets. The results are reported on Table 5 and Table 6. According to Table 5 and Table 6, we can find the proposed Macular still outperform baselines using XLM-R as the backbone. Compared to XLM-R, both CoSDA-ML and Macular performs better on XNLI and QAM since XLM-R does not use code-switch data augmentation. Compared to CoSDA-ML, Macular achieves better performance because the proposed multi-task training framework can share common knowledge across tasks and improve model performance.

Table 5: Performance comparison on the XNLI dataset with XLM-RoBERTa-base [10] as backbone. "AVG" means the average accuracy of all languages. The highest scores per category are in bold. Results of ⋆ are taken from [14].

Method	en	ar	bg	de	el	es	fr	hi
XLM-R⋆	77.7	67.7	72.0	71.7	70.2	72.6	72.7	64.9
CoSDA-ML	84.3	73.9	78.1	77.9	77.1	79.4	78.3	72.4
Macular	**84.9**	**75.0**	**79.5**	**78.5**	**77.6**	**80.0**	**79.8**	**72.8**
Method	ru	sw	th	tr	ur	vi	zh	AVG
XLM-R⋆	70.2	60.7	67.4	69.0	61.0	71.0	69.5	69.1
CoSDA-ML	76.6	**66.7**	73.8	73.9	68.6	76.0	75.8	75.5
Macular	**77.8**	66.3	**74.5**	**74.5**	**69.2**	**77.0**	**76.0**	**76.2**

Table 6: Performance comparison on the QAM dataset with XLM-RoBERTa-base [10] as backbone. "AVG" means the average accuracy of all languages. The highest scores per category are in bold. Results of ⋆ are taken from [25].

Method	en	de	fr	AVG
XLM-R⋆	69.3	**68.1**	67.8	68.4
CoSDA-ML	68.9	67.3	68.2	68.1
Macular	**70.0**	**68.1**	**68.7**	**68.9**

It demonstrates that the proposed Macular is a general framework, which does not rely on a specific backbone.

5 CONCLUSION

In this paper, we explore a novel perspective to tackle the challenging cross-lingual NLU tasks when no parallel corpora are available. Towards this end, we propose a multi-task adversarial framework, namely Macular, which brings mutual performance improvement on both source and target languages. The proposed Macular includes a multi-task module and a task-specific module to infer both the common knowledge across tasks and unique task characteristics. More specifically, we combine a task adversarial loss with task losses defined on both English corpus and corresponding augmented data obtained by code-switch to train the multi-task module for task-shared representation learning. In the task-specific module, we propose to combine language adversarial loss, consistency loss and task loss on source language and augmented data to capture task-specific information. Extensive experiments are conducted on four public datasets including paraphrase identification, natural language understanding, question answering matching, and query advertisement matching. Experimental results show that the proposed Macular outperforms state-of-the-art baselines on all four tasks over multiple languages.

ACKNOWLEDGEMENT

This work is supported in part by the US National Science Foundation under grant NSF IIS-1747614 and IIS-2141037. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Heike Adel, Anton Bryl, David Weiss, and Aliaksei Severyn. 2018. Adversarial neural networks for cross-lingual sequence tagging. *arXiv preprint arXiv:1808.04736* (2018).

[2] Wasi Uddin Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. *arXiv preprint arXiv:2106.02134* (2021).

[3] M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2020. UXLA: A robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. *arXiv preprint arXiv:2004.13240* (2020).

[4] Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518* (2020).

[5] Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje F Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. *arXiv preprint arXiv:2106.02300* (2021).

[6] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2018. Multi-source cross-lingual model transfer: Learning what to share. *arXiv preprint arXiv:1810.03552* (2018).

[7] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL* 6 (2018), 557–570.

[8] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML'08*. 160–167.

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR* 12, ARTICLE (2011), 2493–2537.

[10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Kunbo Ding, Weijie Liu, Yuejian Fang, Weiquan Mao, Zhe Zhao, Tao Zhu, Haoyan Liu, Rong Tian, and Yiren Chen. 2022. A Simple and Effective Method to Improve Zero-Shot Cross-Lingual Transfer Learning. *arXiv preprint arXiv:2210.09934* (2022).

[13] Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. OpenPrompt: An Open-source Framework for Prompt-learning. *arXiv preprint arXiv:2111.01998* (2021).

[14] Xin Luna Dong, Yaxin Zhu, Zuohui Fu, Dongkuan Xu, and Gerard de Melo. 2021. Data Augmentation with Adversarial Training for Cross-Lingual NLI. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5158–5167.

[15] Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231* (2021).

[16] Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2019. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. *arXiv preprint arXiv:1908.07742* (2019).

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[18] Milan Gritta and Ignacio Iacobacci. 2021. Xeroalign: Zero-shot cross-lingual transformer alignment. *arXiv preprint arXiv:2105.02472* (2021).

[19] Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5588–5599.

[20] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587* (2016).

[21] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*. PMLR, 4411–4421.

[22] Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. *arXiv preprint arXiv:2104.08645* (2021).

[23] Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. *arXiv preprint arXiv:1909.00153* (2019).

[24] Chia-Hsuan Lee and Hung-Yi Lee. 2019. Cross-lingual transfer learning for question answering. *arXiv preprint arXiv:1907.06042* (2019).

[25] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401* (2020).

[26] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742* (2017).

[27] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586* (2021).

[28] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. (2015).

[29] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504* (2019).

[30] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742.

[31] Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. *arXiv preprint arXiv:1906.05407* (2019).

[32] Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2020. Multilingual bert post-pretraining alignment. *arXiv preprint arXiv:2010.12547* (2020).

[33] Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. 2020. REA: Robust cross-lingual entity alignment between knowledge graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2175–2184.

[34] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402* (2020).

[35] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *JAIR* 65 (2019), 569–631.

[36] Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6949–6956.

[37] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 567–572.

[38] Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 231–235.

[39] Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. 2019. Cross-lingual word embeddings. *Synthesis Lectures on Human Language Technologies* 12, 2 (2019), 1–132.

[40] Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv preprint arXiv:1809.02306* (2018).

[41] Haoyu Wang, Fenglong Ma, Yaqing Wang, and Jing Gao. 2021. Knowledge-Guided Paraphrase Identification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 843–853.

[42] Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. 2021. Adversarial Domain Adaptation for Cross-lingual Information Retrieval with Multilingual BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3498–3502.

[43] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[44] Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. *arXiv preprint arXiv:1809.03633* (2018).

[45] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934* (2020).

[46] Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. *arXiv preprint arXiv:2106.08226* (2021).

[47] Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*. 437–448.