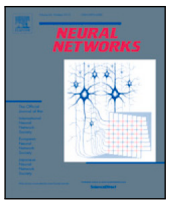
Neural Networks 173 (2024) 106217

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Multi-level multilingual semantic alignment for zero-shot cross-lingual transfer learning

Anchun Gui, Han Xiao *

School of Informatics, Xiamen University, China

ARTICLE INFO

Keywords:
Multilingual semantic alignment
Cross-lingual transfer learning

ABSTRACT

Recently, cross-lingual transfer learning has attracted extensive attention from both academia and industry. Previous studies usually focus only on the single-level alignment (e.g., word-level, sentence-level), based on pre-trained language models. However, it leads to suboptimal performance in downstream tasks of the low-resource language due to the missing correlation of hierarchical semantic information (e.g., sentence-to-word, word-to-word). Therefore, in this paper, we propose a novel multi-level alignment framework, which hierarchically learns the semantic correlation between multiple levels by leveraging well-designed alignment training tasks. In addition, we devise an attention-based fusion mechanism (AFM) to infuse semantic information from high levels. Extensive experiments on mainstream cross-lingual tasks (e.g., text classification, paraphrase identification, and named entity recognition) demonstrate the effectiveness of our proposed method, and also show that our model achieves state-of-the-art performance across various benchmarks compared to other strong baselines.

1. Introduction

In recent years, large-scale pre-trained language models (Chi et al., 2021; Conneau et al., 2020; Devlin et al., 2019) have achieved remarkable success in natural language processing (NLP), even surpassing human performance in specific downstream tasks, e.g., text classification, summarization, and question answering. However, these models are primarily pre-trained on rich-resource languages (e.g., English, Chinese), which leads to unsatisfactory performance on downstream tasks in low-resource languages (e.g., Bulgarian, Hindi). Considering the scarcity of low-resource training corpus and the high cost of pre-training in practice, it is difficult to train a specific model from scratch for each low-resource language. Therefore, this presents a challenge that how to transfer the capability of the model in high-resource languages to low-resource downstream tasks?

To address these concerns, cross-lingual transfer learning is proposed (Artetxe & Schwenk, 2019; Cao et al., 2020; Ding et al., 2022; Eronen et al., 2022; Gritta et al., 2022; Keung et al., 2019; Khurana et al., 2023; Lee et al., 2023; Pan et al., 2021; Sabet et al., 2020; Wang et al., 2023; Xian et al., 2022). Generally speaking, its goal is to align the semantic space between the source language (i.e., the high-resource language) and the target language (i.e., the low-resource language), thereby improving the generalization ability of the model on low-resource languages. At present, the research community has proposed a range of approaches. For example, Artetxe and Schwenk (2019) align word embeddings between the source and target languages through synonym pairs, or bilingual dictionaries (Chaudhary et al., 2020). In addition, some efforts further seek to align the semantic features of parallel sentence pairs to enhance the generalization ability of the model on the target language (Feng et al., 2022; Pan et al., 2021). Despite these single-level alignment methods are straightforward and easily implementable, they fail to yield satisfactory performance on the target language due to the missing correlation of hierarchical semantic information.

To demonstrate this issue formally, we investigate a sentence-level alignment model, LaBSE (Feng et al., 2022), which enhances the cross-lingual transferability of the model by aligning representations between the sentence pairs. As an example in Fig. 1 shows, LaBSE only narrows the representation space between the source language (i.e., "Machine learning is useful" in English) and the target language (i.e., "Maschinelles Lernen ist nützlich" in German) at the sentence-level, while ignoring the semantic relationships between sentence to word and between word to word. In Fig. 2(a), we visualize the correlation heatmap between the source and target sentences and words, where we can observe that, in the single-level alignment, the correlations between all elements are weak except for the parallel sentence pairs. Meanwhile, our multi-level alignment method enhances the correlation between different elements such as sentence-to-word and word-to-word, which further boosts the model performance on

* Corresponding author.
E-mail address: bookman@xmu.edu.cn (H. Xiao).

https://doi.org/10.1016/j.neunet.2024.106217
Received 23 October 2022; Received in revised form 9 October 2023; Accepted 26 February 2024
Available online 27 February 2024

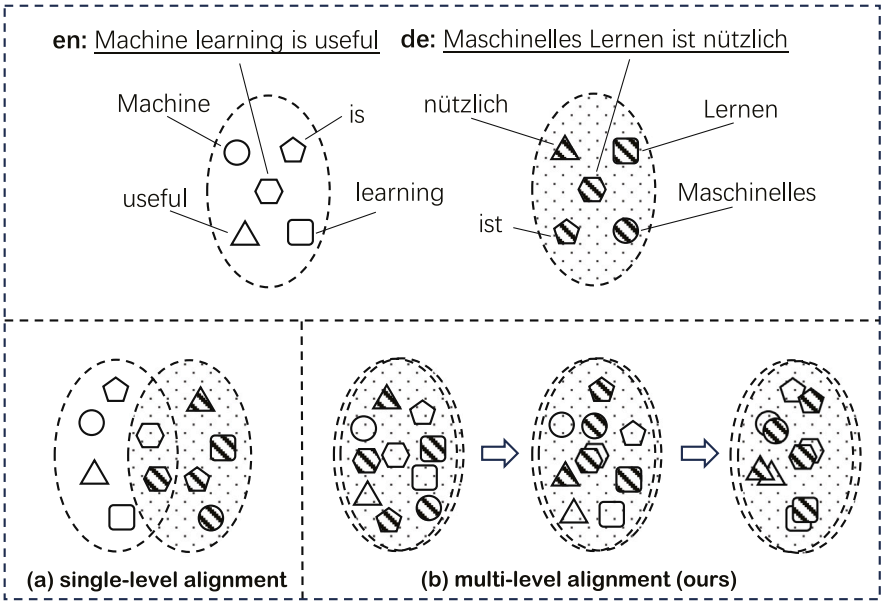


Fig. 1. An example illustrates the comparison between the single-level alignment and our proposed multi-level alignment. (a) The sentence-level alignment approach focuses only on pulling the semantic space between the source and target sentences closer, while ignoring the semantic relationship (e.g., sentence-to-word, word-to-word). (b) In contrast, our proposed multi-level alignment paradigm first narrows the semantic space of two languages holistically, then reduces the representation distance between parallel sentence pairs, and finally aligns the semantic representation between parallel word pairs, respectively.

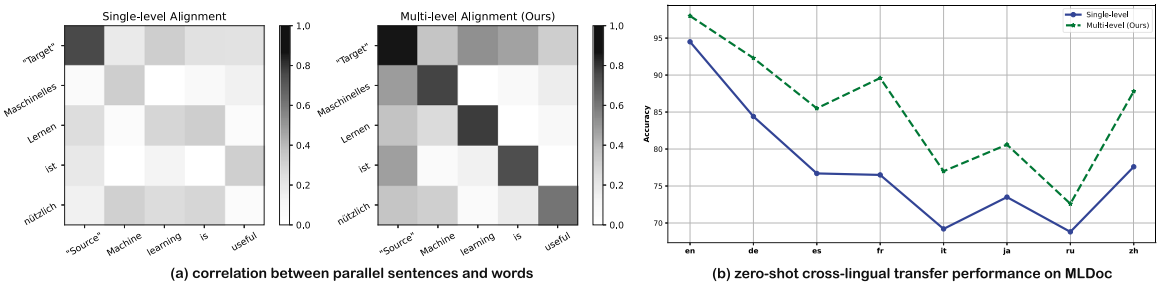


Fig. 2. (a) We visualize the correlation between sentences and words from the source and target languages. Here, "Source" and "Target" represent the source and target sentences, respectively, and darker color means higher correlation. It can be observed that our proposed multi-level alignment method not only enhances the correlation between parallel sentence/word pairs, but also improves the correlation between the source/target sentence-to-word (i.e., diagonal, first column, and first row). (b) We compare the performance between our proposed multi-level alignment model and single-level alignment model (LaBSE) on MLDoc dataset, where "en" is the source language and the others are the target languages. Besides, y-axis represents the accuracy of the model on MLDoc, where our multi-level alignment model outperforms LaBSE by a large margin.

the target language, as shown in Fig. 2(b). Therefore, we believe that the improvement of hierarchical semantic correlation is crucial in cross-lingual transfer learning, and here are some intuitive insights: (i) sentence-level semantics facilitate the analysis of the grammatical structure and lexical comprehension of the language (Siddhant et al., 2020); (ii) word-level semantics are beneficial in dissecting the sentence structure at a fine-grained perspective (Mao et al., 2021).

Motivated by the major concerns in the single-level alignment method, we propose a novel multi-level hierarchical alignment paradigm. Specifically, our method consists of three alignment levels (i.e., language-, sentence-, and word-level), where the alignment process is performed from coarse-grain to fine-grain gradually. First, we introduce adversarial learning (Goodfellow et al., 2014; Keung et al., 2019) to narrow the semantic space of two languages by aligning language-agnostic features. Second, the *inter-sentence* and *intra-sentence* contrastive learning frameworks are leveraged to reduce the representation distance between parallel sentence pairs from the source and target languages. Last, we propose two word-level alignment tasks (i.e., *masked language modeling* (MLM) (Devlin et al., 2019) and *word distribution alignment*) to align the semantic representation between parallel word pairs. In addition, to further strengthen the correlation between different levels of semantics, we devise an *attention-based fusion mechanism* (AFM) module to integrate semantic information from higher levels. The detailed overview of our model is shown in Fig. 4.

To verify our proposed method, we evaluate the model on three challenging multilingual benchmarks: MLDoc (Schwenk & Li, 2018), PAWS-X (Yang et al., 2019), and PANX (Pan et al., 2017). Under the setting of zero-shot cross-lingual transfer learning, our model achieves state-of-the-art results compared to other strong baselines. To summarize, the contributions of this paper are as follows:

- We analyze the weaknesses of the single-level alignment approach and demonstrate the significance of the correlation of hierarchical semantic information in zero-shot cross-lingual transfer learning.
- We propose a novel multi-level, hierarchical alignment paradigm, which significantly enhances the cross-lingual transfer capacity of pre-trained language models by strengthening the semantic correlations between diverse levels (e.g., sentence-to-word).
- Our proposed model obtains state-of-the-art performance on three mainstream benchmarks, and extensive ablation experiments demonstrate the effectiveness of the proposed alignment tasks.

The organization of this paper is as follows. First, we survey the related work of zero-shot cross-lingual transfer learning (Section 2). Then, we introduce our model architecture and proposed alignment

tasks (Section 3). Next, we conduct the experiments to verify and analyze our method (Section 4 & Section 5). Last, concluding remarks are in the final section (Section 6).

2. Related work

In this section, we mainly survey the progress of *multilingual semantic alignment* and *zero-shot cross-lingual transfer learning* in recent years. In addition, we also investigate the advance of *attention mechanism*, which inspires the design of our AFM module.

2.1. Multilingual semantic alignment

Before the advent of large pre-trained multilingual models, to empower the cross-lingual transfer capability, a typical approach is to learn a linear transformation matrix by leveraging the bilingual vocabulary, which projects the source word embeddings into the semantic space of a target language (Mikolov et al., 2013; Xing et al., 2015). Ideally, there is a one-to-one mapping between words in the source and target languages. However, due to the complexity of different languages in practice, this mapping is often difficult to hold. Moreover, this approach cannot handle the semantic information at the sentence-level (Artetxe & Schwenk, 2019).

Recently, with the emergence of pre-trained multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), cross-lingual transfer learning has a new training paradigm. For example, based on these pre-trained multilingual models, some studies attempt to align context-based word representation from the source and target languages (Cao et al., 2020; Mao et al., 2021), thereby further improving the performance of pre-trained multilingual models on the target language. Moreover, there are some studies that explore infusing additional word-level semantic information into the model (Hakimi Parizi & Cook, 2020; Ouyang et al., 2021). In general, these methods can be regarded as the single-level alignment method.

There is another line to boost the cross-lingual transfer capacity of pre-trained multilingual models by aligning parallel sentence pairs (Pan et al., 2021; Wei et al., 2020). For instance, Wei et al. (2020) introduce contrastive learning to narrow the source/target sentence representations from the parallel corpora. Considering the scarcity of parallel corpus especially on low-resource languages, a new framework without parallel corpus is proposed based on instance-weighting strategy (Li et al., 2021). This method leverages the instance-weighted gradient information to update the model parameters. Moreover, inspired by the information-theoretic framework, Chi et al. (2021) propose a new pre-training task for cross-lingual learning from the contrastive learning perspective. Finally, some researchers introduce adversarial learning to enhance the generalization and robustness on the target language (Dong et al., 2020; Huang et al., 2021; Keung et al., 2019).

2.2. Zero-shot cross-lingual transfer learning

To evaluate the cross-lingual transfer capability, zero-shot transfer learning is a widely used setting (Ding et al., 2022; Eronen et al., 2022; Gritta et al., 2022; Gritta & Iacobacci, 2021; Huang et al., 2021; Ji et al., 2020; Keung et al., 2020; Wang et al., 2019; Wu et al., 2022; Xu et al., 2022). Compared to vanilla transfer learning, zero-shot transfer learning is a more challenging setting. As shown in Fig. 3, vanilla transfer learning requires fine-tuning for each target language before inference. However, on one hand, the target languages are usually low-resource and unable to provide task-specific training data (Artetxe & Schwenk, 2019; Ding et al., 2022; Huang et al., 2021; Ji et al., 2020). On the other hand, this paradigm also requires more computational resources and storage space, when dealing with many target languages. Therefore, we expect that the pre-trained models have good generalization and robustness so that they can be directly applied to the downstream tasks in various target languages, with just one fine-tuning on the source training data. This shall be a great challenge for pre-trained language models. Therefore, in this paper, our goal is to design a multi-level alignment framework that enhances the generalization and robustness of the pre-trained model on various target languages.

Downstream tasks. In this paper, we mainly evaluate our model on three fundamental tasks in NLP: text classification, paraphrase recognition, and named entity recognition (NER). (i) Text classification refers to the process of categorizing text data into predefined classes or topics (Minaee et al., 2021). This can be useful for tasks such as sentiment analysis, spam detection, etc (Zhang et al., 2018). (ii) Paraphrase identification is the task of determining whether two text segments express the same meaning or semantic content, even if they use different words or phrasings (Yin & Schütze, 2015). This capability is crucial for numerous NLP applications, such as information retrieval, machine translation, and question-answering (Lan & Xu, 2018). (iii) NER involves the identification and classification of specific textual entities, such as person names, organizations, or locations (Marrero et al., 2013). Through NER, the model can extract vital information, enhance semantic understanding, and provide context-aware solutions, thereby fostering more intelligent and user-friendly language technologies (Liu et al., 2022). Despite some progress made in these research fields before, these tasks remain challenging problems in zero-shot cross-lingual transfer learning setting.

2.3. Attention mechanism

Attention mechanism is a powerful technology in deep learning, which enables the model to selectively focus on specific parts of the input sequence when making predictions. In other words, attention mechanism can be viewed as an automatic weight allocation strategy. Based on this idea, a series of advanced models have been proposed in recent years, such as Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-3 (Brown et al., 2020). For example, BERT equipped with self-attention mechanism achieves superior (even beyond human) performance across various natural language understanding tasks, e.g., paraphrase identification, natural language inference, and sentiment analysis. Meanwhile, the generative models based cross-attention mechanism also obtain tremendous success in natural language generative tasks (Brown et al., 2020; Lewis et al., 2020). Therefore, attention mechanism is introduced into our AFM module to integrate diverse levels of semantic features.

3. Methodology

3.1. Model architecture

Our model architecture is shown in Fig. 4, which consists of three alignment levels (i.e., language-, sentence-, and word-level) from coarse-grain to fine-grain. In the first language-level, the pre-trained multilingual model, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), is viewed as a feature generator, and then an additional discriminator is introduced to form adversarial learning paradigm (Section 3.2). In the second sentence-level, the sentence representation is enhanced by attention fusion mechanism (AFM), which is correlated with the language-level features (Section 3.3). And, two contrastive learning tasks are employed in this level: inter- and intra-sentence alignment (Section 3.4). In the last word-level, the representations, which integrate the semantic information of the language- and sentence-level through AFM module, are used to perform two final word-level training tasks: masked language modeling (MLM, Devlin et al., 2019) and word distribution alignment (Section 3.5).

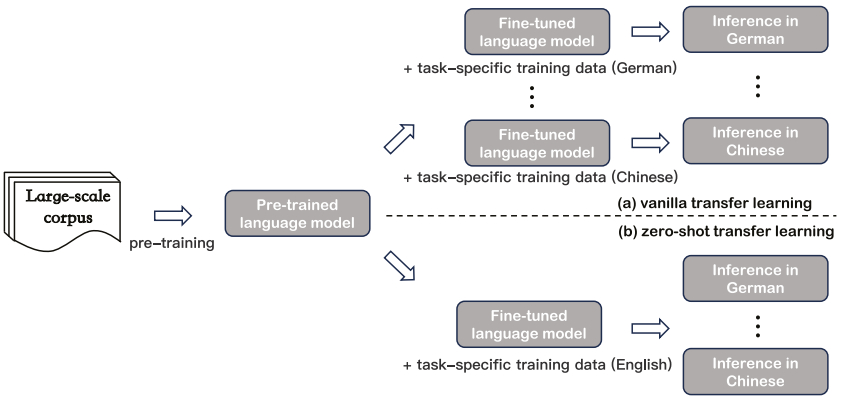


Fig. 3. Comparison between vanilla and zero-shot cross-lingual transfer learning. (a) In vanilla cross-lingual transfer learning, if we intend to adapt the pre-trained model to the specific target languages (e.g., German, Chinese), we need to fine-tune a model for each target language on the corresponding training data. (b) In zero-shot cross-lingual transfer learning, the pre-trained model is fine-tuned on the source training data (i.e., English) only once, then directly applied to the downstream task in different target languages. Therefore, this task requires that the pre-trained model should have good generalization and robustness on various target languages.

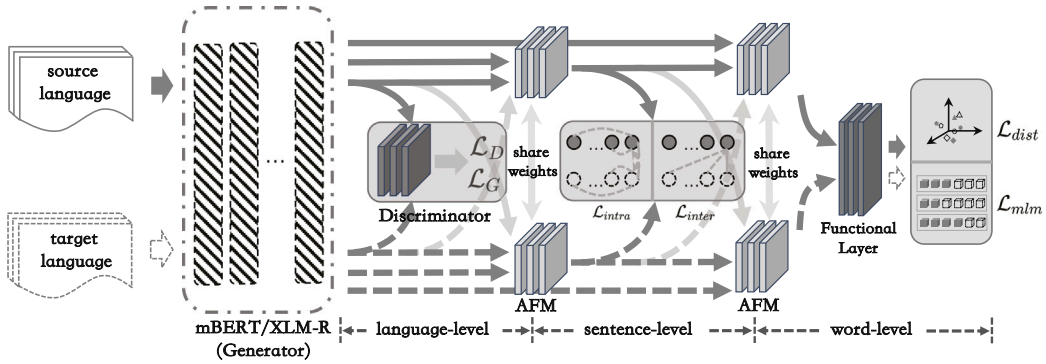


Fig. 4. Model architecture. Our model consists of three alignment levels, which are implemented as follows: (i) We first regard the mBERT/XLM-R (Conneau et al., 2020; Devlin et al., 2019) as a feature generator, and then an extra discriminator is introduced to form adversarial learning paradigm at the language-level. (ii) The sentence representation, which is concatenated with the corresponding target language representation, is fed into the first shared AFM module to obtain rich language-specific semantic information. (iii) Two contrastive learning tasks are employed at the sentence-level: inter- and intra-sentence alignment. (iv) The second shared AFM module and other functional layers (e.g., layer regularization, embedding layer, and softmax function) are leveraged to generate the word representation for the last two word-level alignment training tasks.

3.2. Adversarial learning in language-level

In this level, we take advantage of adversarial learning to align the semantic spaces of the source and target languages. Specifically, we treat the pre-trained multilingual model as a feature generator. Then, a discriminator is introduced to distinguish whether the input sentence is from the source or target language, under the adversarial learning framework (Goodfellow et al., 2014; Keung et al., 2019). Therefore, the objective of the language-level semantic alignment $\mathcal{L}_{\text{LANG}}$ is defined as:

$$\min_{f_G}\max_{f_D}\ \mathbb{E}_{\boldsymbol{x}^{\phi}\sim\mathcal{S}}\left[\log[\sigma(f_D(f_G(\boldsymbol{x}^{\phi})))]\right]+\mathbb{E}_{\boldsymbol{x}^{\psi}\sim\mathcal{T}}\left[\log[1-\sigma(f_D(f_G(\boldsymbol{x}^{\psi})))]\right] \quad (1)$$

where we assume that $\boldsymbol{x}^{\phi}$ and $\boldsymbol{x}^{\psi}$ represent the parallel sentence pairs from the source and target languages, respectively. f_G is a feature generator, and f_D is a discriminator. Besides, $\mathcal{S}$ and $\mathcal{T}$ refer to the source and target sentence sets, separately. $\sigma(\cdot)$ refers to the sigmoid function, whose output could be regarded as the classification probability.

3.3. Attention-based fusion mechanism

To enhance the correlation with higher-level semantics, we design a cross-lingual semantic fusion module based on attention mechanism. To be specific, given the source sentence representation $\boldsymbol{h}_s^{\phi}$, it shall take the source (ϕ) and target (ψ) language representations $\boldsymbol{h}_l^{\phi}$ and $\boldsymbol{h}_l^{\psi}$ into consideration. Since the information of language-level (involving the high-level semantic features) could be a good guidance for the sentence-level alignment tasks. Therefore, we can obtain the new source sentence representation $\boldsymbol{h}_s^{\phi'}$ by leveraging the AFM module:

$$\boldsymbol{h}_s^{\phi'} = \text{AFM}(\boldsymbol{h}_s^{\phi}, \boldsymbol{h}_l^{\phi}, \boldsymbol{h}_l^{\psi}) \quad (2)$$

In addition, the similar operation is also conducted to obtain the new target sentence representation $\boldsymbol{h}_s^{\psi'}$:

$$\boldsymbol{h}_s^{\psi'} = \text{AFM}(\boldsymbol{h}_s^{\psi}, \boldsymbol{h}_l^{\phi}, \boldsymbol{h}_l^{\psi}) \quad (3)$$

where AFM module is defined as follows (take $\boldsymbol{h}_s^{\phi'}$ for example):

$$\alpha_i = \frac{\exp\left(\frac{\boldsymbol{h}_l^i \cdot \boldsymbol{h}_s^{\phi}}{\tau}\right)}{\sum_{j\in\{\phi,\psi\}}\exp\left(\frac{\boldsymbol{h}_l^j \cdot \boldsymbol{h}_s^{\phi}}{\tau}\right)} \quad (4)$$

$$\boldsymbol{h}_s^{*} = \sum_{i\in\{\phi,\psi\}} \alpha_i \boldsymbol{h}_l^i \quad (5)$$

$$\boldsymbol{h}_s^{\phi'} = \sigma(\boldsymbol{W}[\boldsymbol{h}_s^{\phi}; \boldsymbol{h}_s^{*}] + \boldsymbol{b}) \quad (6)$$

where τ is a hyper-parameter, and $\boldsymbol{W}, \boldsymbol{b}$ are trainable weights. Finally, the new source/target sentence representation $\boldsymbol{h}_s^{\phi'}/\boldsymbol{h}_s^{\psi'}$ is applied into the following alignment tasks.

3.4. Contrastive learning in sentence-level

To align the semantic representations between parallel sentence pairs from the source and target languages, we propose two sentence-level contrastive learning tasks: *inter-sentence* and *intra-sentence* alignment. In the inter-sentence task, the parallel sentence pairs are treated as positive instances, while non-parallel sentence pairs are negative instances within a batch. Considering the interchangeability of the source and target languages, we can calculate the contrastive learning loss $\mathcal{L}_{\text{inter}}$ between inter-sentence from the source and target perspectives, respectively. Therefore, the loss $\mathcal{L}_{\text{inter}}$ is formulated to measure the semantic similarity by:

$$\mathcal{L}_{\text{inter}} = -\sum_i \log\left(\frac{\exp(\boldsymbol{h}^{\phi'}_{s(i)} \cdot \boldsymbol{h}^{\psi'}_{s(i)})}{\sum_j \exp(\boldsymbol{h}^{\phi'}_{s(i)} \cdot \boldsymbol{h}^{\psi'}_{s(j)})}\right) - \sum_j \log\left(\frac{\exp(\boldsymbol{h}^{\phi'}_{s(j)} \cdot \boldsymbol{h}^{\psi'}_{s(j)})}{\sum_i \exp(\boldsymbol{h}^{\phi'}_{s(i)} \cdot \boldsymbol{h}^{\psi'}_{s(j)})}\right) \tag{7}$$

where $\boldsymbol{h}^{\phi'}_{s(i)}/\boldsymbol{h}^{\psi'}_{s(j)}$ represents the hidden representations of the i/j-th source/target sentences, which are encoded by the first shared AFM module, as shown in Fig. 4. Here, we could align the semantic features of the sentence-level by maximizing the similarity between parallel sentence pairs.

In the intra-sentence task, we limit contrastive learning into the source or target language scope, where one sentence is used as a positive sample and the other sentences within the same batch serve as negative samples. To this end, we construct the similarity-based intra-sentence contrastive loss objective $\mathcal{L}_{\text{intra}}$, intuitively, which will force the sentence representation to be more distinguishing within the same language (Mao et al., 2021):

$$\mathcal{L}_{\text{intra}} = -\sum_i \sum_j \log\cos\left\{\frac{\pi}{2}\left(\frac{\exp(\boldsymbol{h}^{\phi'}_{s(i)} \cdot \boldsymbol{h}^{\phi'}_{s(j)})}{\sum_k \exp(\boldsymbol{h}^{\phi'}_{s(i)} \cdot \boldsymbol{h}^{\phi'}_{s(k)})} - \frac{\exp(\boldsymbol{h}^{\psi'}_{s(i)} \cdot \boldsymbol{h}^{\psi'}_{s(j)})}{\sum_k \exp(\boldsymbol{h}^{\psi'}_{s(i)} \cdot \boldsymbol{h}^{\psi'}_{s(k)})}\right)\right\} \tag{8}$$

where the similarity between each sentence pair within a batch will be enhanced by optimizing the objective $\mathcal{L}_{\text{intra}}$. Finally, we can obtain the total sentence-level loss function $\mathcal{L}_{\text{SENT}}$:

$$\mathcal{L}_{\text{SENT}} = \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}} \tag{9}$$

3.5. Generative tasks in word-level

To learn more fine-grained semantic information, we propose two word-level training tasks: *masked language modeling* (MLM, Devlin et al., 2019) and *word distribution alignment*. First, let $\boldsymbol{x} = [x_1, x_2, \ldots, x_{\text{mask}}, \ldots, x_n]$ be the input, where x_{mask} refers to a randomly masked token. Given the source/target language input $\boldsymbol{x}^{\phi}/\boldsymbol{x}^{\psi}$, we can obtain its word representation $\boldsymbol{h}^{\phi}_w/\boldsymbol{h}^{\psi}_w$ from the output of the pre-trained multilingual model. Then, the new source/target word representation $\boldsymbol{h}^{\phi'}_w/\boldsymbol{h}^{\psi'}_w$ is fused with the sentence-level semantic information using AFM module (as shown in Fig. 4):

$$\boldsymbol{h}^{\phi'}_w = \text{AFM}(\boldsymbol{h}^{\phi}_w, \boldsymbol{h}^{\phi'}_s, \boldsymbol{h}^{\psi'}_s) \tag{10}$$

$$\boldsymbol{h}^{\psi'}_w = \text{AFM}(\boldsymbol{h}^{\psi}_w, \boldsymbol{h}^{\psi'}_s, \boldsymbol{h}^{\psi'}_s) \tag{11}$$

Next, the masked tokens are predicted by using the new obtained word representations. Here, the loss objective of MLM is defined as:

$$\mathcal{L}_{\text{mlm}} = \sum_{x_{\text{mask}}} -\log p(x_{\text{mask}}|\boldsymbol{h}^{\phi'}_w) - \log p(x_{\text{mask}}|\boldsymbol{h}^{\psi'}_w) \tag{12}$$

To further enhance the word-level alignment effect, the word distributions from parallel sentence pairs (i.e., $\boldsymbol{x}^{\phi}$ and $\boldsymbol{x}^{\psi}$) are aligned. Intuitively, although the words in $\boldsymbol{x}^{\phi}$ and $\boldsymbol{x}^{\psi}$ are from different languages, the sentences composed of them express the same meaning. Therefore, we expect that the word distribution in the source sentence should hold a high similarity to that in corresponding target sentence.

Table 1
Statistics for the three cross-lingual datasets. `#Train` refers to the number of the source training samples, while `#Valid` and `#Test` indicate the number of the validation and test samples for each target language, respectively.

Dataset	#Language	#Class	#Train	#Valid	#Test	Metric
MLDoc	8	4	10 000	1000	4000	Acc.
PAWS-X	7	2	49 401	2000	2000	Acc.
PANX	7	7	20 000	10 000	10 000	F1

Here, we employ KL-divergence to measure the diversity between two word distributions:

$$\mathcal{L}_{\text{dist}} = \sum_{\phi\in\mathcal{S},\psi\in\mathcal{T}} \left(\text{KL}\left(p_{\text{true}}(\boldsymbol{x}^{\phi}) \parallel q_{\text{pred}}(\boldsymbol{x}^{\psi})\right) + \text{KL}\left(p_{\text{true}}(\boldsymbol{x}^{\psi}) \parallel q_{\text{pred}}(\boldsymbol{x}^{\phi})\right)\right) \tag{13}$$

where $p_{\text{true}}(\cdot)$ and $q_{\text{pred}}(\cdot)$ represent the ground-truth word distribution and predicted word distribution, respectively. For the ith token $x^{\phi/\psi}_i$ in the sentence $\boldsymbol{x}^{\phi/\psi}$, the $p_{\text{true}}(\cdot)$ is obtained by the following formula: $p_{\text{true}}(x^{\phi/\psi}_i) = |x^{\phi/\psi}_i| \ / \|\boldsymbol{x}^{\phi/\psi}\|$, where $\|\boldsymbol{x}^{\phi/\psi}\|$ indices the number of tokens in $\boldsymbol{x}^{\phi/\psi}$, and $|x^{\phi/\psi}_i|$ is the frequency of $x^{\phi/\psi}_i$ in $\boldsymbol{x}^{\phi/\psi}$. The calculation of the latter $q_{\text{pred}}(\cdot)$ is derived from the new word representations $\boldsymbol{h}^{\phi'}_w$ and $\boldsymbol{h}^{\psi'}_w$. In addition, since the vocabulary of the multilingual model is relatively large, it may lead to the probability of many words being 0, which is not conducive to optimization. Therefore, we utilize Laplace smoothing to alleviate this issue. Finally, the loss of word-level alignment tasks is obtained by:

$$\mathcal{L}_{\text{WORD}} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{dist}} \tag{14}$$

In summary, the overall loss is a combination of the losses from the three levels, i.e., language-, sentence- and word-level:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{LANG}} + \beta \cdot \mathcal{L}_{\text{SENT}} + \gamma \cdot \mathcal{L}_{\text{WORD}} \tag{15}$$

where λ, β and γ are the coefficient weights of the losses that can be learned using an adaptive optimization strategy (Cipolla et al., 2018) during the training process.

4. Experiments

In this section, we conduct extensive experiments on three mainstream cross-lingual datasets MLDoc (Schwenk & Li, 2018), PAWS-X (Yang et al., 2019), and PANX (Pan et al., 2017) to verify the effectiveness of our proposed method. Statistics for the three datasets are shown in Table 1.

4.1. Setup

Model details. The overview of our model is shown in Fig. 4, where the shape of input $\boldsymbol{x}^{\phi}/\boldsymbol{x}^{\psi}$ is $batch\_size \times seqlen$. For the language representation $\boldsymbol{h}^{\phi/\psi}_l$, it is a special token encoded by the generator and its initial embedding vector is obtained by averaging all word vectors in the corresponding language. Following the previous researches (Devlin et al., 2019; Keung et al., 2019; Mao et al., 2021), we take the context representation of `[CLS]` token as the sentence representation $\boldsymbol{h}^{\phi/\psi}_s$. The dimension of both $\boldsymbol{h}^{\phi/\psi}_l$ and $\boldsymbol{h}^{\phi/\psi}_s$ is $batch\_size \times d_{model}$, while the word representation $\boldsymbol{h}^{\phi/\psi}_w$ is $batch\_size \times seqlen \times d_{model}$. Besides, the discriminator is a multi-layer perceptron, and AFM modules are built based on Eq. (2)–(6). In our experiments, $batch\_size = 96$, $seqlen = 128$, and $d_{model} = 768$.

Table 2
In zero-shot setting, we report the accuracy on MLDoc test sets with different target languages. We compare our proposed models (i.e., Multi-Level alignment) with the previous baselines based on Transformer and non-Transformer architectures, respectively. Here, our results are averaged across five training runs and pass the significant paired t-test (i.e., $p < 0.05$).

Model	en	de	es	fr	it	ja	ru	zh	Avg.
Non-Transformer Architecture									
Multi-CCA (Schwenk & Li, 2018)	92.2	81.2	72.5	72.4	69.4	67.6	60.8	74.7	73.8
LASER (Artetxe & Schwenk, 2019)	89.9	84.8	77.3	77.9	69.4	60.3	67.8	71.9	74.9
Hakimi Parizi and Cook (2020)	–	80.2	78.5	86.5	72.9	74.6	53.3	69.6	73.6
Transformer Architecture									
mBERT (Keung et al., 2019)[a]	94.2	79.8	72.1	73.5	63.7	72.8	73.7	76.0	75.7
mBERT+Adv (Keung et al., 2019)	–	88.1	80.8	85.7	72.3	76.8	77.4	84.7	80.8
T-LASER (Li & Mak, 2020)	–	84.6	73.8	74.9	70.5	–	–	–	–
mBERT+IW (Li et al., 2021)	–	87.6	75.3	75.2	71.2	72.6	69.4	82.7	76.2
Mao et al. (2021)	–	88.8	80.8	85.1	74.3	–	–	–	–
LaBSE (Feng et al., 2022)	94.5	84.4	76.7	76.5	69.2	73.5	68.8	77.6	77.7
XLM-R (Nishikawa et al., 2021)[a]	94.4	86.7	81.5	84.9	73.4	78.5	71.3	85.2	82.0
mBERT+ML (Ours)	97.8	90.1	84.9	89.2	74.8	76.9	**78.1**	86.5	84.8
XLM-R+ML (Ours)	**98.0**	**92.3**	**85.4**	**89.4**	**77.0**	**80.6**	72.5	**87.7**	**85.4**

Bold indices the best result.
[a] Indicates that the results are reported in the corresponding paper.

Implementation details. In the pre-training stage, CCMatrix-v1 dataset[1] is used as our training corpora, which contains 4.5 billion parallel pairs and covers 576 languages. Considering our limited computation resource, we only sample a subset as our pre-training parallel corpus. Specifically, we choose English as the source language and 16 other languages (i.e., Arabic, Bulgarian, German, Greek, Spanish, French, Hindi, Italian, Japanese, Russian, Swahili, Thai, Turkish, Urdu, Vietnamese, and Chinese) as the target languages. These target languages cover the entire downstream tasks in our experiments. We leverage the open-source HuggingFace library[2] to implement our proposed method. To begin with, we initialize the backbone model exactly as the official model weights, while additional modules are randomly initialized. The optimizer is Adam, and different learning rates are set for different tasks. For example, since the discriminator is trained from scratch, its learning rate is initially assigned with 8×10^{-4}, which gradually decays as the training progresses. Meanwhile, for the generator initialized by the pre-trained model, we set a relatively small learning rate. After several trials, 5×10^{-6} is chosen for better performance. In addition, to balance the losses between different tasks, we employ an adaptive loss optimization strategy (Cipolla et al., 2018). As shown in Eq. (15), each loss is equipped with a trainable weight that can be learned adaptively.

Zero-shot setting. In zero-shot setting, we only fine-tune our model on the training data of the source language (i.e., English), and then directly applied to other target language tasks without any additional fine-tuning procedures. Note that, under zero-shot cross-lingual transfer setting, the model merely accesses the source data labels, while the labels of target data are not available. Therefore, this is a highly challenging task that could adequately evaluate the cross-lingual transfer capacity of the model.

4.2. Cross-lingual news text classification

MLDoc[3] is a broadly used dataset to evaluate the capability of the multilingual model in cross-lingual transfer classification (Hakimi Parizi & Cook, 2020; Mao et al., 2021). This dataset includes four types of news documents written in eight languages (i.e., English, Chinese, French, German, Italian, Japanese, Russian, and Spanish). More concretely, this dataset is a four-class (i.e., Corporate, Economics, Government, and Markets) classification task. Additionally, this dataset is a class-balanced subset from the Reuters News RCV1 and RCV2 datasets.[4] In Table 2, we compare our proposed method with the previous strong baselines, and our model achieves new state-of-the-art performance. Moreover, from the comparison between our models based on different backbones (i.e., mBERT and XLM-R), XLM-R is a better backbone than mBERT in most target languages except for Russian. We conjecture that this may be due to the poor quality of Russian corpus in the pre-training phase of XLM-R. Note that although the accuracy in English is relatively high (which is reasonable since the pre-trained model is fine-tuned on labeled English dataset), we are primarily interested in the performance of non-English languages.

4.3. Cross-lingual paraphrase identification

Paraphrase identification is also a widely used downstream task in NLP. This task aims to determine whether two given sentences have the same meaning (Yang et al., 2019). This requires that the model has strong semantic parsing ability and could grasp the essence of semantic expression without being affected by perturbations (e.g., synonym substitution). Here, PAWS-X dataset[5] is used in our experiments, and it contains seven different languages: English, French, Spanish, German, Chinese, Japanese, and Korean. The experimental setting is zero-shot cross-lingual transfer learning, and the results are reported in Table 3. Compared to the previous baselines, our best model (XLM-R+ML) achieves state-of-the-art results on target languages and consistently outperforms the mBERT-based models. This demonstrates that we proposed multi-level alignment framework could significantly improve the zero-shot cross-lingual transfer capability of the model.

4.4. Cross-lingual named entity recognition

Unlike the text classification and paraphrase identification tasks, the named entity recognition (NER) is a more challenging task. Since NER involves the recognition and classification of specific textual entities, such as person names, organizations, or locations, it requires the model to understand multiple languages at a more fine-grained word-level. We evaluate the models on PANX (Pan et al., 2017), which is extracted from Wikipedia. The results are shown in Table 4, where we can see that our proposed model consistently outperforms the other baselines. Remarkably, the model obtains better performance on those target languages (e.g., German, French) that have high similarity to the source language (i.e., English). In comparison, the performance still needs to be further improved on those target languages (e.g., Arabic, Russian) that differ significantly from the source language.

[1] https://opus.nlpl.eu/CCMatrix.php
[2] https://github.com/huggingface/transformers
[3] https://github.com/facebookresearch/MLDoc
[4] https://trec.nist.gov/data/reuters/reuters.html
[5] https://github.com/google-research-datasets/paws

Table 3
The accuracy results of the zero-shot cross-lingual transfer learning on PAWS-X. The results are averaged across five training rounds and pass the significant paired t-test (i.e., $p < 0.05$).

Model	fr	es	de	zh	ja	ko	Avg.
Yang et al. (2019)	85.2	86.0	82.2	75.8	70.5	71.7	78.5
Ahmad et al. (2019)	87.0	87.4	85.7	77.0	73.0	69.6	79.9
Artetxe et al. (2020)	85.2	85.5	81.6	72.5	–	–	–
Glavaš and Vulić (2021)	87.1	–	85.8	78.7	–	–	–
mBERT (Yang et al., 2019)†	85.2	86.0	82.2	75.8	70.5	71.7	78.5
XLM-R[a]	86.4	86.2	83.2	79.1	78.4	76.2	81.5
mBERT+ML (Ours)	87.4	87.5	85.3	79.1	74.6	74.2	81.3
XLM-R+ML (Ours)	**89.0**	**88.1**	**87.4**	**81.5**	**79.6**	**79.8**	**84.2**

The highest scores are in **bold**.
[a] Refers to our reproduced results.

Table 4
F1 scores of the zero-shot cross-lingual named entity recognition task are evaluated on PANX. The results are averaged across five training rounds and pass the significant paired t-test (i.e., $p < 0.05$).

Model	en	ar	de	es	fr	ru	vi	Avg.
mBERT (Devlin et al., 2019)	83.1	46.9	73.5	72.4	76.2	63.2	65.4	68.7
XLM (Conneau & Lample, 2019)	82.4	50.6	74.8	69.8	77.2	61.0	66.5	68.9
XLM-R (Conneau et al., 2020)	83.3	50.1	73.3	74.5	77.0	63.2	64.7	69.4
InfoXLM (Chi et al., 2021)	83.7	54.9	74.2	71.1	77.4	64.5	65.9	70.2
mBERT+ML (Ours)	84.0	48.3	74.8	74.2	78.7	64.2	67.2	70.2
XLM-R+ML (Ours)	**84.2**	**55.2**	**75.5**	**75.3**	**79.2**	**65.6**	**67.5**	**71.4**

The highest scores are in **bold**.

Table 5
Ablation experiments on MLDoc dataset.

Model	en	de	es	fr	it	ja	ru	zh	Avg.
XLM-R+ML	98.0	92.3	85.4	89.4	77.0	80.6	72.5	87.7	85.4
w/o AFM	97.0	91.7	82.7	84.5	75.2	78.2	70.7	85.5	83.2
w/o lang-level	96.2	91.2	81.6	87.5	73.8	77.7	70.8	85.2	83.0
w/o sent-level	96.5	87.7	83.5	87.7	74.2	78.9	71.4	86.5	83.3
w/o intra-sent	97.6	89.5	85.4	87.1	76.8	78.2	72.9	86.6	84.3
w/o inter-sent	96.9	88.3	84.7	88.3	75.4	79.8	71.4	87.8	84.1
w/o word-level	96.6	88.8	84.7	88.2	76.2	76.3	71.4	84.4	83.3
w/o word-mlm	96.4	90.2	83.8	89.0	76.3	79.4	71.3	85.1	84.0
w/o word-dist	96.8	89.6	84.7	88.9	76.4	78.4	71.8	86.2	84.1
XLM-R[a]	95.1	87.2	81.7	85.6	74.5	78.3	71.2	86.1	82.5

[a] Indicates the model is re-trained on more corpus compared with the original model.

Table 6
Ablation experiments on PANX dataset.

Model	en	ar	de	es	fr	ru	vi	Avg.
XLM-R+ML	84.2	55.2	75.5	75.3	79.2	65.6	67.5	71.4
w/o AFM	80.1	49.2	71.3	73.0	71.5	60.3	65.4	67.3
w/o lang-level	79.8	51.6	69.6	70.8	73.8	62.8	65.1	67.6
w/o sent-level	77.4	50.6	68.7	69.1	72.9	60.4	62.4	65.9
w/o intra-sent	77.7	51.5	67.0	70.4	73.3	61.3	63.7	66.4
w/o inter-sent	77.9	51.8	68.1	69.8	72.8	62.3	63.8	66.6
w/o word-level	78.7	52.2	68.2	72.1	71.5	61.5	63.4	66.8
w/o word-mlm	79.8	50.9	69.8	73.5	72.3	60.3	64.8	67.3
w/o word-dist	79.4	51.0	68.9	74.0	72.4	61.8	64.3	67.4
XLM-R*	76.2	48.7	68.6	69.6	70.3	61.2	62.7	65.3

5. Analysis

5.1. The effect of proposed methods

To verify the effectiveness of our proposed three levels alignment tasks and AFM module, we conduct ablation experiments to explore their effects. In the experiments, XLM-R is used as our backbone model, and the performance of various variant models is evaluated on MLDoc and PANX datasets.

Effect of adversarial learning. We remove adversarial learning and maintain the other alignment tasks unchanged. Meanwhile, we keep the model structure and hyper-parameters settings consistent with the previous ones. This experiment is denoted as **w/o lang-level**. As can be seen in Tables 5 and 6, compared to the standard baseline (i.e., XLM-R+ML), the model performance has a significant decline without the adversarial learning objective. Intuitively, on one hand, the introduced adversarial task could effectively align the language-agnostic features across various languages. On the other hand, the lack of semantic information at the language-level also hinders the progress of the latter two levels of alignment tasks. Therefore, it reveals that adversarial learning we proposed is effective and necessary.

Effect of contrastive learning. We eliminate the whole sentence-level alignment tasks (denoted **w/o sent-level**), intra-sentence task (denoted **w/o intra-sent**), and inter-sentence task (denoted **w/o inter-sent**) to explore their effects, respectively. The comparison results are reported in Tables 5 and 6, where the model suffers performance degradation on most target languages without the second alignment level, especially in Table 6 of NER task.

Effect of generative tasks. Similar to the prior experimental setups, the total word-level tasks (denoted **w/o word-level**), MLM task (denoted **w/o word-mlm**), and word distribution alignment task (denoted **w/o word-dist**) are removed, respectively. From the comparison results in Tables 5 and 6, we can see that the performance on downstream task decreases compared to the standard baseline, unsurprisingly. Therefore, it demonstrates that the fine-grained semantic information from the word-level is beneficial to the alignment of different semantic spaces.

Effect of AFM. As for the AFM module (denoted **w/o AFM**), the model performance drops $\sim 2\%$ without this procedure. In summary, through the ablation experiments in Tables 5 and 6, the rationality and necessity of each alignment level and AFM module are verified.

5.2. Complexity analysis

The time complexity. Although our proposed alignment framework consists of three levels, its implementation efficiency is actually

Table 7
Case study on MLDoc dataset, where the collection of categories includes: "Corporate", "Economics", "Government", and "Markets". Here, "Single-level" and "Multi-level" represent the predicted results of the single-level alignment model and our proposed multi-level alignment model, respectively. Here, the blue words mean that they may have a strong relationship with the predicted results of our model in the incorrect predictions.

Target	Sentence	Label	Single-level	Multi-level (Ours)
de	Die Exportpreise Australien sind im Juli gegenüber dem Vormonat um 0,1 Prozent gesunken. Im Jahresabstand hätten sich die Güter sogar um 7,6 Prozent verbilligt, teilte das Statistische Amt am Montag in Canberra mit.	Economics	Government ✗	Economics ✓
	Die Schweizerische Nationalbank (SNB) lässt den Diskontsatz unverändert bei einem Prozent. Dies sagte ein SNB-Sprecher am Donnerstag auf Anfrage.	Economics	Government ✗	Government ✗
zh	芝加哥期貨交易所(CBOT)稻米期貨週五收盤走堅,因週末前少量補空買盤介入.交易商表示,亞洲現貨米市尤其是泰國米疲軟,繼續抑制此間米市漲勢.各月稻米每cwt收盤上漲2至9美分不等,11月期約上揚7-1／2美分成爲10.14-1／2美元.稻米期貨成交量估計250張,週四爲545張.	Markets	Economics ✗	Markets ✓
	台灣央行官員週二表示,決定今日不釋出100億台幣郵儲金利息.央行有關官員在接受路透社記者電話訪問時,做出上述表示.央行有關官員並未說明爲何取消該項釋金計劃,但據接近業務局高層官員人士表示,〃因故〃而臨時取消該項釋金行動,但目前尚不知何時會釋出該款項.郵匯局資金管理人員先前表示,該局已接獲央行通知釋出100億台幣郵儲金利息,但之後又接獲央行通知取消釋金計劃.	Markets	Economics ✗	Government ✗

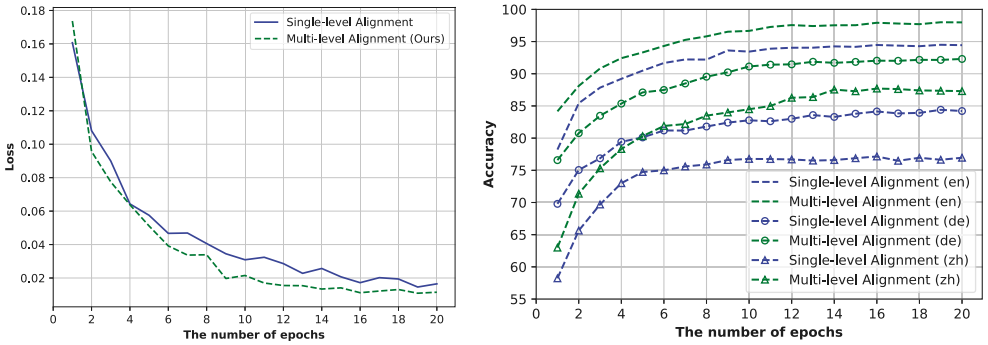


Fig. 5. Left: Comparison of the convergence between our proposed multi-level alignment method and the single-level alignment method in the training process. **Right:** Comparison of the performance on the MLDoc's test dataset between our multi-level model and the single-level model across the source language (i.e., en) and the target languages (i.e., de and zh), with the increase of the training epoch.

the same with the baseline (i.e., the single-level alignment approach). In practice, the tasks of these three levels can be performed simultaneously, as shown in Eq. (15), where the losses of the language-, sentence-, and word-level are added up as a holistic optimization objective during the training process. As for the AFM module, it only involves three operations: softmax, summation, and the linear transformation. Here, the time complexity of each operation is $\mathcal{O}(n)$, where n is the dimension of the input sequence. Therefore, the overall time complexity of AFM module is $\mathcal{O}(n)$ as well.

The training convergence & model generalization. Despite that the time complexity of our approach is the same as the baseline, here we empirically demonstrate that our method has significant advantages over the baseline in terms of both the training convergence and model's generalization capacity. Specifically, based on MLDoc dataset, we fine-tune our model (XLM-R+ML) and the baseline model (LaBSE) separately on the training set of the source language (i.e., en), and then record the model losses as the training epoch increases, as shown in the left of Fig. 5. Meanwhile, we also save the checkpoints of models at each epoch, and then directly evaluate the current model's performance on the test datasets of the source and target (i.e., de and zh) languages. Note that this evaluation on the target languages are completely conducted in the zero-shot cross-lingual setting, which is

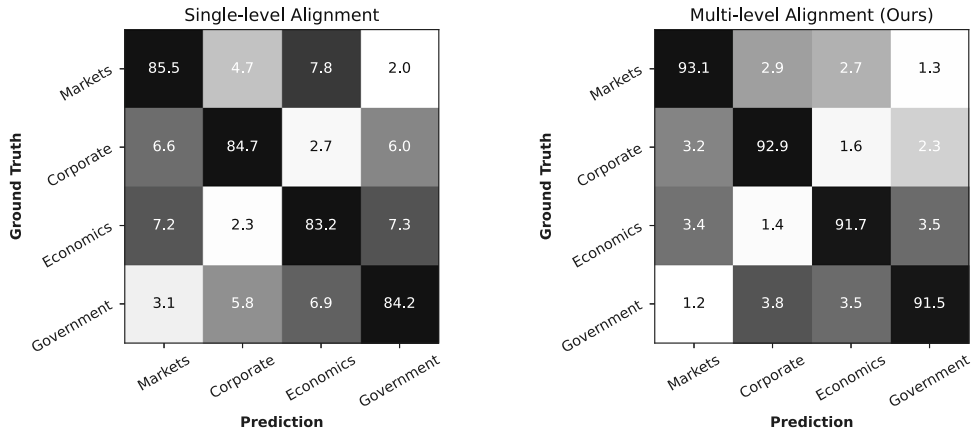


Fig. 6. Comparison of the confusion matrix between our proposed multi-level alignment method and the single-level alignment method on the MLDoc's test dataset from the target language (German), under the zero-shot cross-lingual setting.

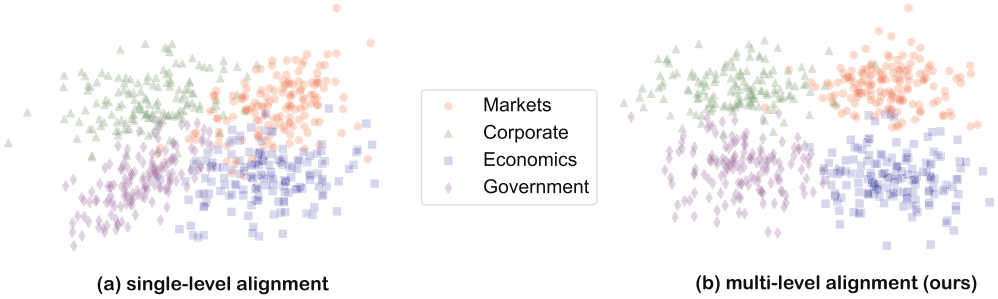


Fig. 7. Comparison of the clustering effect based on our proposed multi-level alignment method and the single-level alignment method, on the MLDoc's test set from the target language (German).

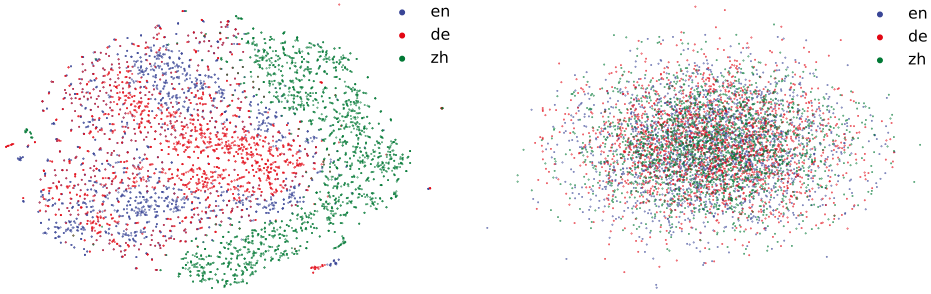


Fig. 8. The t-SNE visualization of different semantic spaces are embedded by the original pre-trained model (**left**), and our proposed multi-level alignment model (**right**).

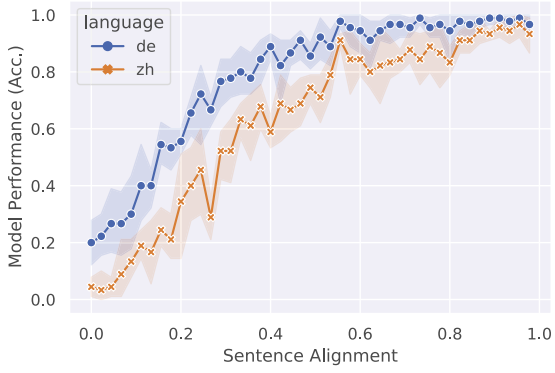


Fig. 9. The effect of the sentence-level semantic alignment on the transfer performance across languages. The x-axis represents the degree of alignment of the source sentence (English) and target sentences (German and Chinese), and the y-axis refers to the model performance on MLDoc dataset. The results of x/y-axis are normalized, and the closer representation of the source and target sentences means better cross-lingual performance.

a highly challenging setting but also an effective way to measure the generalization on other target languages. The experimental results are shown in Fig. 5, where we can observe that: (i) Regarding the training convergence, our model is able to achieve faster convergence efficiency with the help of joint optimization of multi-tasks. For example, under the same training epoch, our model has lower loss on the training dataset and higher performance on the test dataset. (ii) Regarding the generalization capacity, compared to the baseline method, our model can obtain significant improvements not only in the source language but also in the target languages, especially on Chinese target dataset. This demonstrates that our proposed multi-level alignment framework can effectively improve the generalization of the model.

5.3. Case study

In this subsection, we perform a detailed case study to analyze the performance differences between our proposed method and the baseline, and attempt to provide some reasonable explanations. Specifically, in the zero-shot cross-lingual setting, we directly apply the fine-tuned model on the MLDoc training set from the source language

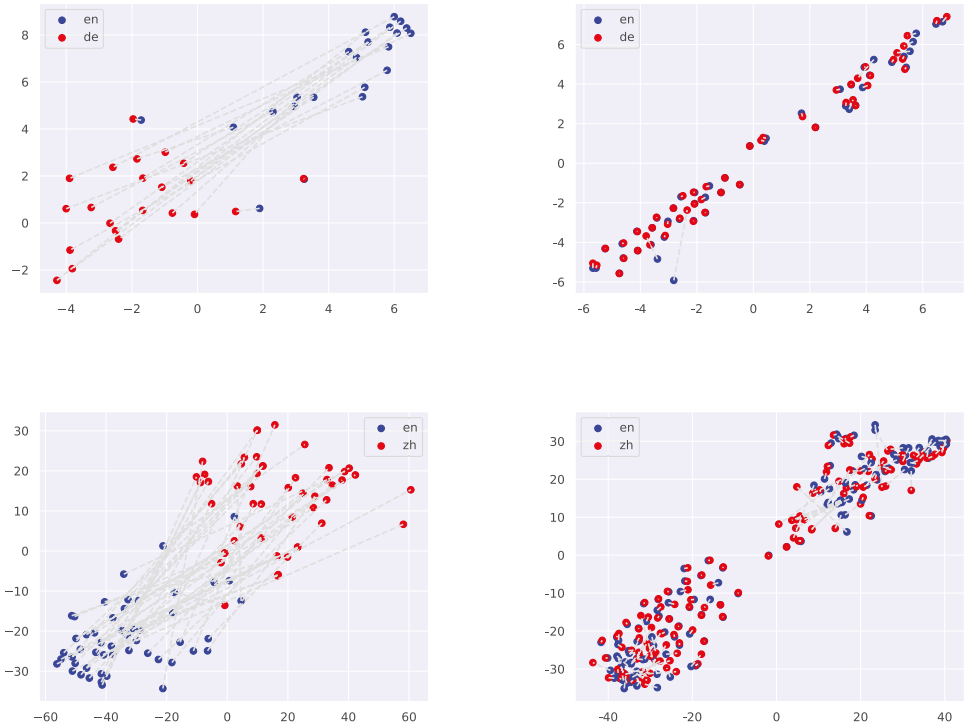


Fig. 10. The t-SNE projection of parallel word pairs (connected by the light gray dashed line) from the source and target languages. The left side is a visualization of the original model, while the right side is a visualization of our multi-level alignment model.

(i.e., English) to the test set of the target language (i.e., German) without additional fine-tuning process. Here, LaBSE and XLM-R+ML are leveraged as baselines of the single- and multi-level methods, respectively, and other configurations are consistent with Section 4.1. In fact, from Table 2, we can see that the accuracies of our model and the baseline on the test set of German are 92.3% and 84.4%, respectively. To further understand the model's classification performance on different categories (there are four categories in the MLDoc: *Markets*, *Corporate*, *Economics*, and *Government*), we plot the confusion matrices of different models in Fig. 6. It can be observed that after training with our proposed multi-level alignment framework, the model not only significantly improves the accuracy in each category (refer to the comparison of elements on the diagonal in Fig. 6), but also greatly reduces the percentage of misclassification. For example, the baseline misclassifies *Markets* as *Economics* with a share of 7.8%, while our approach decreases it to 2.7%, which substantially enhances the accuracy of *Markets*.

In addition, to more intuitively understand the superiority of our approach compared to the baseline, we visualize the feature distribution of samples in the test set from German after encoding by different models, and the visualization results are shown in Fig. 7. We can observe that the baseline indeed clusters the samples into four classes, but it is obvious that there are more overlapping regions between diverse classes, which is particularly unfavorable for the model to learn classification boundaries. In contrast, the clustering based on our approach significantly reduces the overlapping regions and has relatively clear decision boundaries, such as *Markets* and *Economics*. It explains why our approach remarkably boosts the model's generalization capacity in the target languages. Finally, we provide some bad cases that are on the classification boundary and misclassified by the model, as shown in Table 7, where we can observe that these misclassified samples often involve multiple categories of information. For example, the last Chinese sentence in Table 7 contains both market and government information, which leads to ambiguous classifications. Therefore, the reason for these misclassifications is somehow attributed to the lack of fine-grained categories in the MLDoc. Nevertheless, our multi-level alignment model still has stronger discriminative capability than the single-level alignment model.

5.4. Visualization and further exploration

We visualize the alignment effect at the language-level and word-level after implementing our alignment tasks, respectively. As for the sentence-level alignment, since the visualization is similar to word-level, we instead explore the impact of similarity between the source and target languages on the cross-lingual transfer performance.

Language-level alignment. To evaluate the alignment effect of our multi-level alignment model at the language-level, we leverage the t-SNE algorithm to project all word embedding vectors of the language vocabulary (e.g., English, German, and Chinese) into a two-dimensional plane, where each color indicates a language-level semantic representation space, as shown in Fig. 8. We can draw that the "semantic cloud space" of different languages are mixed after implementing our proposed alignment tasks.

Sentence-level alignment. Here, the experimental settings are similar to Section 4.1, where the source language is English and the target languages are German and Chinese. Our detailed implementation steps are as follows: we first obtain the sentence similarity across languages by calculating the cosine similarity of corresponding parallel sentence pairs. Then, multiple subsets are selected according to their similarity. Last, for a certain sentence similarity, the transfer capacity of the model is tested on MLDoc dataset (involving German and Chinese). The experimental results are shown in Fig. 9, where we can conclude that the better performance is attained on the target language with the sentence alignment enhancing. Even at the identical degree of sentence alignment, the better results are achieved on the target language (German) that is more similar to the source language (English).

Word-level alignment. To further demonstrate the effect of the proposed word-level alignment, we also utilize the t-SNE algorithm to project the word pairs embedding from different languages (e.g., en-de and en-zh)

into a two-dimensional plane. As depicted in Fig. 10, the word pairs from the source and target languages are significantly aligned after implementing our proposed alignment tasks.

6. Conclusion

In this paper, we propose a multi-level, hierarchical alignment training framework to improve the cross-lingual transfer learning capability of the pre-trained language model. Compared to the preceding single-level alignment method, which lacks the correlation between hierarchical semantic information, our proposed method effectively learns multiple-level semantic knowledge and gradually fuses them through AFM module. To evaluate the effectiveness of our proposed method, we test the model on mainstream cross-lingual benchmarks. The results show that our model outperforms the previous baselines under the same evaluation protocol. In addition, extensive ablation experiments also demonstrate the validity and necessity of each alignment task we proposed.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xiamen University, China

Data availability

Data will be made available on request.

Acknowledgments

This work is funded by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62006201).

References

Ahmad, W., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., & Peng, N. (2019). On difficulties of cross-lingual transfer with order differences: a case study on dependency parsing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 2440–2452). doi:10/gnxxv9.

Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4623–4637). doi:10/ghgndg.

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610, doi:10/gkz62j.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *vol. 33*, *Advances in neural information processing systems* (pp. 1877–1901).

Cao, S., Kitaev, N., & Klein, D. (2020). Multilingual alignment of contextual word representations. In *8th international conference on learning representations*.

Chaudhary, A., Raman, K., Srinivasan, K., & Chen, J. (2020). DICT-MLM: Improved multilingual pre-training using bilingual dictionaries. arXiv:2010.12566, [cs].

Chi, Z., Dong, L., Wei, F., Yang, N., Singhal, S., Wang, W., Song, X., Mao, X.-L., Huang, H., & Zhou, M. (2021). InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 3576–3588). doi:10/gnhzx2.

Cipolla, R., Gal, Y., & Kendall, A. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7482–7491). doi:10/gf9ns2.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). doi:10/gm72p7.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Vol. 32*, In *Advances in neural information processing systems*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). doi:10/ggbwf6.

Ding, K., Liu, W., Fang, Y., Mao, W., Zhao, Z., Zhu, T., Liu, H., Tian, R., & Chen, Y. (2022). A simple and effective method to improve zero-shot cross-lingual transfer learning. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4372–4380).

Dong, X., Zhu, Y., Zhang, Y., Fu, Z., Xu, D., Yang, S., & de Melo, G. (2020). Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1541–1544). doi:10/gn3bk4.

Eronen, J., Ptaszynski, M., Masui, F., Arata, M., Leliwa, G., & Wroczynski, M. (2022). Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management*, *59*(4), Article 102981. http://dx.doi.org/10.1016/j.ipm.2022.102981.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 878–891). http://dx.doi.org/10.18653/v1/2022.acl-long.62.

Glavaš, G., & Vulić, I. (2021). Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume* (pp. 3090–3104). doi:10/gn3bmp.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Vol. 27*, In *Advances in neural information processing systems*.

Gritta, M., Hu, R., & Iacobacci, I. (2022). CrossAligner & Co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the association for computational linguistics: ACL 2022* (pp. 4048–4061). http://dx.doi.org/10.18653/v1/2022.findings-acl.319.

Gritta, M., & Iacobacci, I. (2021). XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 371–381). doi:10/gnv3m2.

Hakimi Parizi, A., & Cook, P. (2020). Joint training for learning cross-lingual embeddings with sub-word information without parallel corpora. In *Proceedings of the ninth joint conference on lexical and computational semantics* (pp. 39–49).

Huang, K.-H., Ahmad, W., Peng, N., & Chang, K.-W. (2021). Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1684–1697).

Ji, B., Zhang, Z., Duan, X., Zhang, M., Chen, B., & Luo, W. (2020). Cross-lingual pre-training based transfer for zero-shot neural machine translation. *Vol. 34*, In *Proceedings of the AAAI Conference on Artificial Intelligence* (01), (pp. 115–122). http://dx.doi.org/10.1609/aaai.v34i01.5341.

Keung, P., Lu, Y., & Bhardwaj, V. (2019). Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 1355–1360). doi:10/gngnh8.

Keung, P., Lu, Y., Salazar, J., & Bhardwaj, V. (2020). Don't use english dev: on the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 549–554). doi:10/gm3njt.

Khurana, S., Dawalatabad, N., Laurent, A., Vicente, L., Gimeno, P., Mingote, V., & Glass, J. (2023). Improved cross-lingual transfer learning for automatic speech translation. http://dx.doi.org/10.48550/arXiv.2306.00789, arXiv:2306.00789.

Lan, W., & Xu, W. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th international conference on computational linguistics*.

Lee, H., Yoon, H.-W., Kim, J.-H., & Kim, J.-M. (2023). Cross-lingual transfer learning for phrase break prediction with multilingual language model. http://dx.doi.org/10.48550/arXiv.2306.02579, arXiv:2306.02579.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). http://dx.doi.org/10.18653/v1/2020.acl-main.703.

Li, W., & Mak, B. (2020). Transformer based multilingual document embedding model. arXiv:2008.08567, [cs].

Li, I., Sen, P., Zhu, H., Li, Y., & Radev, D. (2021). Improving cross-lingual text classification with zero-shot instance-weighting. In *Proceedings of the 6th workshop on representation learning for NLP (repL4NLP-2021)* (pp. 1–7). doi:10/gm6j45.

Liu, P., Guo, Y., Wang, F., & Li, G. (2022). Chinese named entity recognition: The state of the art. *Neurocomputing*, *473*, 37–53.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692, [cs].

Mao, Z., Gupta, P., Chu, C., Jaggi, M., & Kurohashi, S. (2021). Lightweight cross-lingual sentence representation learning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 2902–2913). doi:10/gm6jgw.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv:1309.4168, [cs].

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40.

Nishikawa, S., Yamada, I., Tsuruoka, Y., & Echizen, I. (2021). A multilingual bag-of-entities model for zero-shot cross-lingual text classification. arXiv:2110.07792, [cs].

Ouyang, X., Wang, S., Pang, C., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 27–38).

Pan, L., Hang, C.-W., Qi, H., Shah, A., Potdar, S., & Yu, M. (2021). Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 210–219). doi:10/gnhsnc.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1946–1958). http://dx.doi.org/10.18653/v1/P17-1178.

Sabet, A., Gupta, P., Cordonnier, J.-B., West, R., & Jaggi, M. (2020). Robust cross-lingual embeddings from parallel sentences. arXiv:1912.12481, [cs].

Schwenk, H., & Li, X. (2018). A corpus for multilingual document classification in eight languages. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Siddhant, A., Johnson, M., Tsai, H., Ari, N., Riesa, J., Bapna, A., Firat, O., & Raman, K. (2020). Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *Vol. 34*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8854–8861). doi:10/gngnhn.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *30*, In *Advances in neural information processing systems*.

Wang, Y., Che, W., Guo, J., Liu, Y., & Liu, T. (2019). Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5721–5727). doi:10/ghm9j5.

Wang, J., Meng, F., Zheng, D., Liang, Y., Li, Z., Qu, J., & Zhou, J. (2023). Towards unifying multi-lingual and cross-lingual summarization. http://dx.doi.org/10.48550/arXiv.2305.09220, arXiv:2305.09220.

Wei, X., Weng, R., Hu, Y., Xing, L., Yu, H., & Luo, W. (2020). On learning universal representations across languages. In *International conference on learning representations*.

Wu, i., Wu, S., Zhang, X., Xiong, D., Chen, S., Zhuang, Z., & Feng, Z. (2022). Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. arXiv:2204.00996, [cs].

Xian, R., Ji, H., & Zhao, H. (2022). Cross-lingual transfer with class-weighted language-invariant representations. In *International conference on learning representations*.

Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1006–1011). doi:10/gnxw6d.

Xu, Q., Baevski, A., & Auli, M. (2022). Simple and effective zero-shot cross-lingual phoneme recognition. In *Interspeech 2022* (pp. 2113–2117). http://dx.doi.org/10.21437/Interspeech.2022-60.

Yang, Y., Zhang, Y., Tar, C., & Baldridge, J. (2019). PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3687–3692). doi:10/gkz654.

Yin, W., & Schütze, H. (2015). Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 901–911).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), Article e1253.