Multi-level Distillation of Semantic Knowledge for Pre-training Multilingual Language Model

Mingqi Li[1], **Fei Ding**[1], **Dan Zhang**[1], **Long Cheng**[1], **Hongxin Hu**[2], **Feng Luo**[1*]
[1]Clemson University, [2]University at Buffalo
{mingqil,feid,dzhang4,lcheng2,luofeng}@clemson.edu
hongxinh@buffalo.edu

Abstract

Pre-trained multilingual language models play an important role in cross-lingual natural language understanding tasks. However, existing methods did not focus on learning the semantic structure of representation, and thus could not optimize their performance. In this paper, we propose Multi-level Multilingual Knowledge Distillation (MMKD), a novel method for improving multilingual language models. Specifically, we employ a teacher-student framework to adopt rich semantic representation knowledge in English BERT. We propose token-, word-, sentence-, and structure-level alignment objectives to encourage multiple levels of consistency between source-target pairs and correlation similarity between teacher and student models. We conduct experiments on cross-lingual evaluation benchmarks including XNLI, PAWS-X, and XQuAD. Experimental results show that MMKD outperforms other baseline models of similar size on XNLI and XQuAD and obtains comparable performance on PAWS-X. Especially, MMKD obtains significant performance gains on low-resource languages.

1 Introduction

Pre-training a large-scale language model and fine-tuning it on downstream tasks has shown great success in natural language processing. Most works (Devlin et al., 2018; Liu et al., 2019; Yang et al., 2019b) focused on English, since it is easy to get a large amount of training data for English. This paradigm has recently emerged as a promising means for cross-lingual tasks. Some multilingual language models (Devlin et al., 2018; Conneau et al., 2019) trained on monolingual data from over 100 languages using Masked Language Modeling (MLM) objective performed surprisingly well without any explicit alignment. On the other hand, XLM (Lample and Conneau, 2019) extended MLM objective to a parallel corpus version - Translation Language Modeling (TLM), and achieved impressive results. This inspired researchers to develop alignment methods using parallel corpora.

Follow-up works of XLM (Yang et al., 2020; Wei et al., 2020; Chi et al., 2020; Ouyang et al., 2020) leveraged various of training objectives to align parallel sentences at different granularity. These works were usually trained using both large amounts of monolingual data and additional parallel corpora, which are time-consuming and require considerable computational resources.

Another line of research (Cao et al., 2020; Pan et al., 2020; Hu et al., 2020a) only used limited parallel data to improve existing pre-trained language models rather than training new models from scratch. They depended on new alignment methods for parallel pairs of words and sentences, which could further achieve performance gains over current state-of-the-art pre-trained language models. However, these approaches neglected vector space properties when aligning across languages, and thus generating sub-optimal results. We hypothesize that a large-scale English corpus can provide more semantic and structural information than most other languages used to train multilingual language models. Moreover, BERT (Devlin et al., 2018), which is trained from a vast amount of English Wikipedia and BooksCorpus (Zhu et al., 2015), can capture this information properly and guide the training procedure of other languages, especially for those with limited resources.

In this work, we employ a teacher-student framework to adopt vector space properties in English, and transfer its rich knowledge to our multilingual language model. We use BERT-base as the teacher model and Multilingual BERT (mBERT; Devlin et al., 2018) as the student model. We propose a Multi-level Multilingual Knowledge Distillation (MMKD) method to align semantically similar sentences in parallel corpora to improve mBERT. Specifically, we propose a Cross-lingual

*Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3097 - 3106
December 7-11, 2022

Word-aware Contrastive Learning (XWCL) to encourage word representation similarity between teacher and student networks. We also adopt TLM objective in the student network to take advantage of corresponding context information in the target languages of masked tokens. We present a new sentence-level alignment objective to imitate English sentence projections from the teacher network. Moreover, we propose a structure-level alignment objective to transfer relationships between sentences in BERT vector space. We conduct experiments on zero-shot cross-lingual natural language understanding tasks, including natural language inference, paraphrase identification, and question answering. Experimental results show that MMKD significantly improves mBERT and outperforms baseline models of similar size. The analysis demonstrates the cross-lingual transferability of MMKD on low-resource languages. MMKD provides a more feasible and effective pre-training procedure that only requires limited training data and fewer computational resources.

2 Related Work

2.1 Multilingual Language Model Pre-training

Several efforts trained multilingual language models with transformer-based architectures and large-scale monolingual corpora across over 100 languages. For instance, Devlin et al. (2018) trained Multilingual BERT (mBERT) on 104 languages with objectives of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), and it performed surprisingly well without any explicit alignment. XLM-R (Conneau et al., 2019) achieved further performance gains by leveraging more training data and a larger model.

One popular family of methods proposed different training objectives to align words or sentences from parallel corpora. XLM (Lample and Conneau, 2019) extended MLM objective to a parallel corpus version - Translation Language Modeling (TLM). Unicoder (Huang et al., 2019) improved the transferability by presenting five pre-training tasks. ALM (Yang et al., 2020) predicted words in the context of code-switching sentences. HICTL (Wei et al., 2020) introduced sentence-level and word-level alignment with contrastive learning. INFOXLM (Chi et al., 2020) proposed cross-lingual contrast (XLCO) to maximize mutual information of sentence pairs. ERNIE-M (Ouyang et al., 2020) presented cross-attention masked language modeling (CAMLM) and back-translation masked modeling (BTMLM) to leverage both parallel and monolingual corpora.

More recently, researchers considered computational resources and time and presented works based on existing multilingual language models. Cao et al. (2020) minimized the similarity between word pairs in parallel sentences in a post-hoc manner. Pan et al. (2020) argued that creating word alignments using FastAlign (Dyer et al., 2013) would suffer from the noise of the toolkit and neglected the contextual information. They proposed Post-Pretraining Alignment (PPA) that combined a different TLM objective and a contrastive learning objective. AMBER (Hu et al., 2020a) presented objectives that encouraged prediction of the corresponding sentence and consistency between attention matrices, and they pre-trained the model with an extremely large batch size of 8,192 for the first 1M steps.

2.2 Knowledge Distillation

Hinton et al. (2015) first introduced knowledge distillation to transfer knowledge to a small model, and it has been widely used for transferring dark knowledge (which refers to information that can tell us how the model tends to generalize) and model compression in Natural Language Processing and Computer Vision. A series of follow-up works achieved gains on multilingual tasks. Sun et al. (2020) enhanced the generalization ability of unsupervised neural machine translation by adding self-knowledge distillation and language branch knowledge distillation. Wang et al. (2020) reduced the distance between monolingual teachers and the multilingual student to predict multilingual label sequences. To the best of our knowledge, Reimers and Gurevych (2020) is the only multilingual language model related work that applied a student model to mimic sentence representations generated from the teacher model. They fed both source and target sentences into the student model to calculate Mean Square Error (MSE) loss with the teacher model's source sentences.

3 Methodology

This section presents the training procedure and introduces our four proposed training objectives. Our goal is to improve multilingual language models by transferring semantic knowledge from English

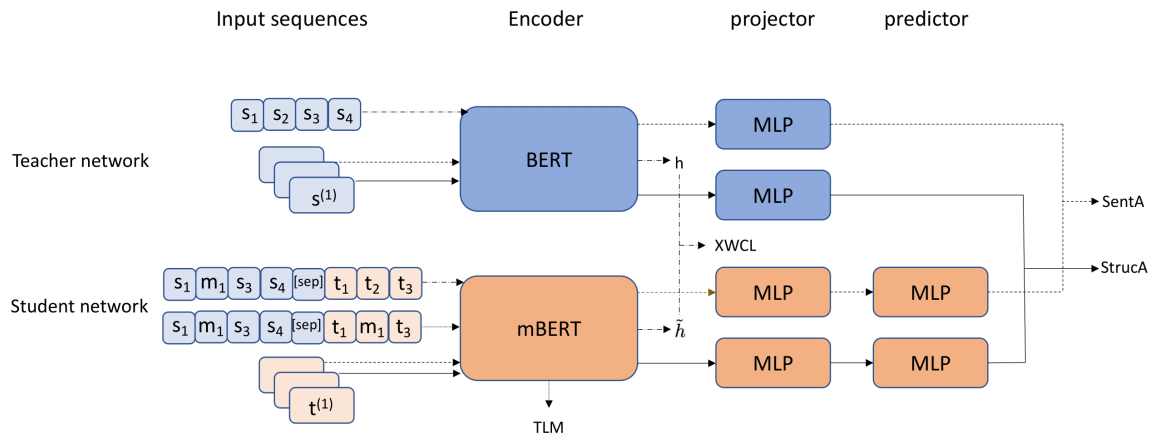


Figure 1: Model Architecture of our proposed Multi-level Multilingual Knowledge Distillation method which combines TLM, XWCL, SentA, and StrucA objectives and is trained in a multi-task manner.

and aligning multi-level information in parallel corpora with limited resources. The general network architecture is illustrated in Figure 1. The student network consists of three components: an encoder, a projector, and a predictor, while the teacher network contains an encoder and a projector.

3.1 Translation Language Modeling

Translation Language Modeling (TLM) objective is an extension of MLM (Lample and Conneau, 2019). Given the concatenation of parallel sentences, TLM objective predicts masks in both source and target sequences. In this way, TLM utilizes context information in the corresponding language, and thus helps the model to learn token-level alignments.

Similar to Devlin et al. (2018), we randomly mask 15% tokens from input sequences and replace them with a [MASK] token 80% of the time, with a random token in vocabulary 10% of the time, and keep them unchanged 10% of the time. The input sequence is denoted as $[s_1, \ldots, s_a, \text{[SEP]}, t_1, \ldots, t_b]$, where a, b are numbers of tokens, and masks exist in both source and target sides. Since the teacher model only involves English, we train TLM objective on the student model.

3.2 Cross-lingual Word-aware Contrastive Learning

Inspired by Su et al. (2021), we propose a cross-lingual version of word-aware contrastive learning (XWCL) objective. The goal of XWCL is to encourage the student model to learn more discriminative representations. Different from Su et al. (2021), our student model produces representations according to the parallel context instead of surrounding monolingual words. Moreover, due to the vocabulary difference in our teacher and student models, we align the representations on the word-level.

Given an English source sequence $s = [s_1, \ldots, s_n]$ and a target sequence $t = [t_1, \ldots, t_m]$, we concatenate them with a special token [SEP] and randomly mask 15% words only from source sequence s following the same mask strategy in Devlin et al. (2018). Then, we feed this masked sequence into the student model and get representation $\tilde{h} = [\tilde{h}_1, \ldots, \tilde{h}_{n+m}]$. Meanwhile, we input the original sequence s into the teacher model and get $h = [h_1, \ldots, h_n]$ as reference. Our proposed XWCL objective learns to minimize the infoNCE loss of the masked tokens:

$$\mathcal{L}_{\text{XWCL}} = -\sum_{i=1}^{n} \log m(s_i) \frac{\exp\left(\text{sim}\left(\tilde{h}_i, h_i\right)/\tau\right)}{\sum_{j=1}^{n} \exp\left(\text{sim}\left(\tilde{h}_i, h_j\right)/\tau\right)}, \quad (1)$$

where τ is a temperature parameter, $\text{sim}(\cdot,\cdot)$ denotes dot product, $m(s_i) = 1$ if s_i is a masked token, otherwise $m(s_i) = 0$. Here we mask the whole word and treat the first token of each mask as the word representation. Consequently, XWCL will make masked representations produced by the student model closer to their corresponding representations in English vector space, and discriminate them from other distinct representations.

3.3 Sentence Alignment

BERT is well-trained with a large-scale English corpus and thus encodes rich semantic knowledge. The goal of our proposed Sentence Alignment (SentA) objective is to capture this semantic information and transfer it to mBERT. Similar to

Grill et al. (2020), we learn representations by instance-level discrimination without negative samples, while we freeze the teacher model rather than updating with an exponential moving average.

Given a sentence pair $(s^{(i)}, t^{(i)})$ in parallel corpora, where $s^{(i)}$ is the i-th sentence from English and $t^{(i)}$ is from a target language, we treat them as two different views and input $s^{(i)}$ into the teacher network and $t^{(i)}$ into the student network separately. We minimize Mean Squared Error loss between teacher projections and student predictions:

$$\mathcal{L}_{\text{SentA}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\bar{q}_\theta \left(g_\theta \left(\tilde{c}^{(i)} \right) \right) - \bar{g}_\xi \left(c^{(i)} \right) \right)^2, \quad (2)$$

where $c^{(i)}$ and $\tilde{c}^{(i)}$ are the [CLS] tokens of last hidden states of i-th sentence in the teacher and student encoders, g defines the projectors with distinct parameters and q defines the predictor, and $\bar{g}$, $\bar{q}$ indicate that they are normalized with L_2 norm. More precisely, we apply 2-layers MLPs to implement projectors and the predictor.

SentA objective will force different languages closer to semantically similar English sentences in the vector space. Meanwhile, the student network can adopt well-trained English vector space properties by imitating corresponding representations in the teacher network.

3.4 Structure Alignment

Transferring relationships between samples plays a crucial role in knowledge distillation. Inspired by Ding et al. (2020), we propose a Structure Alignment (StrucA) objective to learn knowledge correlation.

Given a batch of source-target sentence pairs $((s^{(1)}, t^{(1)}), \ldots, (s^{(\mathcal{B})}, t^{(\mathcal{B})}))$, we feed them into the same teacher-student encoders as calculating SentA objective, while using their own projection and prediction heads. Let $z = [z^{(1)}, \ldots, z^{(\mathcal{B})}]$ and $\tilde{z} = [\tilde{z}^{(1)}, \ldots, \tilde{z}^{(\mathcal{B})}]$ denote teacher projections and student predictions. The proposed objective allows the student network to mimic the vector space structure of the teacher network, which means the correlation between $\tilde{z}$ is similar to z. Specifically, we first normalize z and calculate the similarity matrix:

$$\mathcal{A}_{i,j} = z^{(i)} \cdot z^{(j)}, \tilde{\mathcal{A}}_{i,j} = \tilde{z}^{(i)} \cdot \tilde{z}^{(j)}. \quad (3)$$

Then, the teacher's relational function can be expressed as:

$$\psi \left(z^{(1)}, .., z^{(\mathcal{B})} \right) = \frac{\exp \left(\mathcal{A}_{i,j} / \tau \right)}{\sum_j \exp \left(\mathcal{A}_{i,j} / \tau \right)}. \quad (4)$$

The student network follows the same step, but takes $log\_softmax$ function as a relational function instead. Finally, we employ KL-divergence loss to minimize the difference between two probability distributions:

$$\mathcal{L}_{\text{StrucA}} = \sum_{i=1}^{\mathcal{B}} \text{KLDivLoss} \left(\psi \left(\cdot \right), \tilde{\psi} \left(\cdot \right) \right). \quad (5)$$

After training, the relationship between samples produced by the student network in vector space will be similar to their counterparts in English. In conclusion, StrucA objective learns additional structural information in English vector space.

3.5 Multi-level Multilingual Knowledge Distillation Pre-training

We jointly train these proposed objectives that cover alignments at different granularity, and the final loss would be:

$$\mathcal{L} = \mathcal{L}_{\text{TLM}} + \mathcal{L}_{\text{XWCL}} + \mathcal{L}_{\text{SentA}} + \alpha \mathcal{L}_{\text{StrucA}}, \quad (6)$$

where α is used to balance the weights.

For training, we update the student network by AdamW (Loshchilov and Hutter, 2017) optimizer, while freezing parameters in the teacher model.

In addition, we randomly shuffle the sentence pairs from each parallel datasets, but balance the number of samples from each language within a batch. In other words, our model will consider each language of the same weight during the training procedure.

4 Experiments

This section explains our training details and shows the experimental results on XNLI, PAWS-X and XQuAD. We compare our proposed MMKD with existing works following the setting in Hu et al. (2020b) and conduct ablation studies to prove the effectiveness of each proposed objective.

4.1 Training Details

Training Data We collect the same parallel corpora as previous works (Cao et al., 2020; Pan et al., 2020) for comparison. Specifically, we treat English as source language and download datasets from the OPUS website (Tiedemann, 2012) including (1) low-resource languages: en-hi from IITB (Kunchukuttan et al., 2017) and en-bg from EUbookshop and Europarl (Koehn, 2005) (2)

Model	Vocab size	Layers	Parameters	Ratio	Data
mBERT	119K	12	178M	1.0x	Wikipedia
MONOTRANS	30K	12	110M	0.62x	Wikipedia
PPA	110K	12	172M	0.97x	10.5M parallel data
AMBER	120K	12	172M	0.97x	Wikipedia + 58.5M parallel data
Cao et al. (2020)	119K	12	178M	1.0x	1.8M parallel data
MMKD (this work)	119K	12	179M	1.01x	10.5M parallel data
MMTE	64K	6	192M	1.08x	103 languages in-house parallel data
mT5	250K	12	580M	3.26x	CommonCrawl
XLM-100	200K	12	828M	4.65x	Wikipedia
XLM-R-Large	250K	24	816M	4.58x	CommonCrawl

Table 1: Model size and training data for comparison. Ratio is the parameters' ratio of mBERT. Wikipedia and CommonCrawl are extremely larger than other parallel datasets. For AMBER, we only list the parallel data size of the languages we consider. The numbers of parameters in PPA and AMBER are slightly different from mBERT we use.

high-resource languages: en-ar, en-zh from MultiUN (Eisele and Chen, 2010) and en-fr, en-es, en-de from Europarl. Our training data does not involve any monolingual corpora; however, mBERT is trained from large-scale monolingual corpora. Additionally, we remove extremely short (less than 10 tokens) and long (more than 128 tokens) sentences and prune each dataset to 2M sentence pairs if they contain more than that. Table 1 indicates the size of parallel datasets in our work.

Model Architecture The architecture of teacher and student encoders is the same as BERT-base, which contains 12 layers, 768 hidden states, and 12 attention heads. The student encoder is initialized from mBERT, while the teacher encoder is initialized from BERT-base, and thus they have different vocabulary sizes. Additionally, the student encoder is followed by two different projection and prediction heads, and two projection heads are also on the top of the teacher encoder. These heads consist of randomly initialized 2-layer MLP with 768 hidden dimensions and 128 output dimensions.

Training Setups During the training procedure, we optimize the student network by AdamW with 1e-2 weight decay and schedule the learning rate with a linear decay peaking at 2e-5 after 10% warm-up steps. We set 128 tokens as the maximum length of each sequence and use a batch size of 256. The training procedure takes 3 days for 15 epochs on 8 40GB Nvidia A100 GPUs. For the evaluation procedure, we fine-tune the student encoder for few epochs with a batch size of 32 on English training data, and evaluate on target languages.

4.2 Evaluation Benchmarks

We evaluate our multilingual language model using publicly available cross-lingual natural language understanding benchmarks, including natural language inference, paraphrase identification, and question answering tasks. We conduct all the experiments with a zero-shot setting: we fine-tune the model on English training data and directly test on target languages.

XNLI Conneau et al. (2018) is a widely used cross-lingual sentence classification dataset that extends SNLI/MultiNLI (Bowman et al., 2015; Williams et al., 2018) in fifteen languages. The task is to classify the relationships of two given sentences to entailment, neutral, or contradiction. This dataset provides 2490 dev samples and 5010 test samples in each language. In the zero-shot setting, we fine-tune our student encoder using English MultiNLI, which contains 392,702 sentence pairs. Then, we select the model according to the performance on XNLI English dev set, and test target languages using XNLI test set.

PAWS-X The goal of PAWS-X (Yang et al., 2019a) is to identify whether the two sentences are paraphrases. This dataset translates Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019) evaluation pairs in six languages. We take 49,401 English training pairs in PAWS as training data, and use around 2,000 sentence pairs of each target language from PAWS-X as testing data.

XQuAD Artetxe et al. (2019) requires to return an answer span derived from the paragraph according to the question. Professional translators translate a subset of SQuAD v1.1 (Rajpurkar et al., 2016) development set into ten languages, contain-

Models	en	ar	bg	de	es	fr	hi	zh	avg
Model size similar to mBERT									
mBERT (Devlin et al., 2018)	81.4	62.1	-	70.5	74.3	-	-	63.8	-
mBERT* (Devlin et al., 2018)	80.8	64.3	68.0	70.0	73.5	73.4	58.9	67.8	69.6
MONOTRANS (Artetxe et al., 2019)	81.7	70.6	73.7	73.0	75.4	74.7	65.2	70.3	73.1
Cao et al. (2020)	80.1	-	73.4	73.1	75.5	74.5	-	-	-
PPA (Pan et al., 2020)	82.8	70.3	73.8	74.2	76.7	76.6	66.9	**72.8**	74.3
AMBER (Hu et al., 2020a)	**84.7**	70.2	74.3	74.2	76.9	76.6	66.2	71.6	74.3
MMTE* (Siddhant et al., 2020)	79.6	64.9	70.4	68.2	71.6	69.5	63.5	69.2	69.6
MMKD (this work)	83.2	**71.6**	**75.8**	**76.2**	**78.0**	**77.2**	**69.0**	72.3	**75.4**
Larger models									
mT5-Base (Xue et al., 2020)	84.7	73.3	78.6	77.4	80.3	79.1	70.8	74.1	77.3
XLM-100* (Lample and Conneau, 2019)	82.8	66.0	71.9	72.7	75.5	74.3	62.5	70.2	72.0
XLM-R-Large* (Conneau et al., 2019)	**88.7**	**77.2**	**83.0**	**82.5**	**83.7**	**82.2**	**75.6**	**78.2**	**81.4**

Table 2: Zero-shot cross-lingual classification evaluation results on XNLI. * indicates the results are taken from Hu et al. (2020b). All other results are from original papers.

ing 1,290 question-answering pairs. We use 87,599 training data in SQuAD v1.1 together with 1,190 testing data of each target language in XQuAD.

4.3 Results and Analysis

We report the results across the above evaluation benchmarks. We compare our pre-trained model with mBERT (Devlin et al., 2018) and the following mBERT-based models: (1) Cao et al. (2020); (2) PPA (Pan et al., 2020); (3) AMBER (Hu et al., 2020a). These models adopt BERT-base architecture and are initialized from mBERT. We also compare with MONOTRANS (Artetxe et al., 2019) and MMTE (Siddhant et al., 2020), which contain a similar amount of parameters to mBERT. These models use relatively fewer computational resources than larger models, but still achieve some gains. We also take the results of large models as reference including: (1) mT5 (Xue et al., 2020); (2) XLM-100 (Lample and Conneau, 2019); (3) XLM-R-Large (Conneau et al., 2019). These models are usually costly, but they boost the state-of-the-art on many cross-lingual tasks. There is a resources-performance trade-off in training multilingual language models. Table 1 shows the model size and training data.

XNLI Table 2 presents zero-shot cross-lingual classification accuracy on XNLI. Similar to Cao et al. (2020) and Pan et al. (2020), we evaluate only on the languages used in the pre-training procedure. We first compare with zero-shot results of mBERT to see whether our alignment method improves the existing model. We take the mBERT results from Hu et al. (2020b) to provide a more comprehensive comparison. Our model significantly outperforms mBERT across all the reported languages. We obtain a boost in performance of 5.8% accuracy on average. Moreover, we observe significant improvements on Bulgarian and Hindi which are considered as low-resource languages in our experiments. We only collect 370k en-bg sentence pairs and 895k en-hi sentence pairs. However, they outperform mBERT by 7.8% and 10.1% respectively.

Compared to the models of similar size, we reach a state-of-the-art of 75.4% across eight languages on the zero-shot XNLI benchmark dataset. We significantly outperform these models on ar, bg, de, es, fr, hi and achieve comparable results on en and zh. The results show that we obtain consistent improvements except Chinese compared to PPA (Pan et al., 2020). Additionally, the performance of Latin-based languages is better than non-Latin-based languages in our case. We conclude that our method is more beneficial to languages closed to English, since we adopt English BERT as the teacher model and all the training sentence pairs involve English.

Our method also produces 3.4% gains compared to XLM-100, despite the fact that we have 78% fewer parameters than theirs. Our model is 6% less than XLM-R-Large, which is 4.5 times larger and employs a much more extensive training dataset.

PAWS-X Table 3 reports zero-shot cross-lingual paraphrase identification accuracy on PAWS-X. In this experiment, we evaluate on five high-resource languages involved in our pre-training step. The high resource helps to learn rich information and structured semantic representations; thus, existing models can perform well across these languages. We push mBERT classification accuracy

Models	en	de	es	fr	zh	avg
Model size similar to mBERT						
mBERT* (Devlin et al., 2018)	94.0	85.7	87.4	87.0	77.0	86.2
MONOTRANS (Artetxe et al., 2019)	94.3	86.3	87.6	87.3	79.0	86.9
AMBER (Hu et al., 2020a)	**95.6**	**89.4**	89.2	**90.7**	80.9	**89.2**
MMTE* (Siddhant et al., 2020)	93.1	85.1	87.2	86.9	75.9	85.6
MMKD (this work)	94.7	88.8	**89.7**	88.9	**81.1**	88.6
Larger models						
mT5-Base (Xue et al., 2020)	**95.4**	89.4	89.6	**91.2**	81.1	89.2
XLM-100* (Lample and Conneau, 2019)	94.0	85.9	88.3	87.4	76.5	86.4
XLM-R-Large* (Conneau et al., 2019)	94.7	**89.7**	**90.1**	90.4	**82.3**	**89.4**

Table 3: Zero-shot cross-lingual paraphrase identification evaluation accuracy on PAWS-X. * indicates the results are taken from Hu et al. (2020b). All other results are from original papers.

Models	en	ar	de	es	hi	zh	avg
Model size similar to mBERT							
mBERT* (Devlin et al., 2018)	83.5	61.5	70.6	75.5	59.2	58.0	68.1
MONOTRANS (Artetxe et al., 2019)	82.1	**66.0**	70.6	70.8	61.9	**60.5**	68.7
MMTE* (Siddhant et al., 2020)	80.1	63.2	68.8	72.4	61.3	55.8	66.9
MMKD (this work)	**84.8**	64.0	**73.5**	**76.7**	**62.7**	58.8	**70.1**
Larger models							
mT5-Base (Xue et al., 2020)	84.6	63.8	73.8	74.8	60.3	**66.1**	70.6
XLM-100* (Lample and Conneau, 2019)	74.2	61.4	66.0	68.2	56.6	49.7	62.7
XLM-R* (Conneau et al., 2019)	**86.5**	**68.6**	**80.4**	**82.0**	**76.7**	59.3	**75.6**

Table 4: Zero-shot cross-lingual question answering evaluation F1 score on XQuAD. * indicates the results are taken from Hu et al. (2020b). All other results are from original papers.

from 86.2% to 88.6% with the help of alignment objectives.

MMKD outperforms models of similar size by an accuracy of 1.7%-3% except AMBER. One primary reason is that AMBER is trained with an extremely large batch size that has proven effective by Liu et al. (2019). Another reason is that PAWS-X only consists of high-resource languages, and thus can not demonstrate our model's benefit on low-resource languages. Similar to XNLI, we observe consistent improvements over XLM-100 on the paraphrase identification task. Compared to XLM-R-Large, we bridge the performance gap to 0.8% with limited computational resources.

XQuAD Table 4 shows zero-shot cross-lingual question answering F1 score on XQuAD. Our model obtains the best F1 score on average against other baseline models of similar size. The results on four Latin-based languages significantly outperform other models, while our model produces relatively small gains on Arabic and Chinese. This is consistent with our findings on XNLI.

For larger models, we outperform XLM-100 across all evaluation languages and achieve comparable results to mT5-Base whose parameters are much more than ours.

4.4 Ablation Study

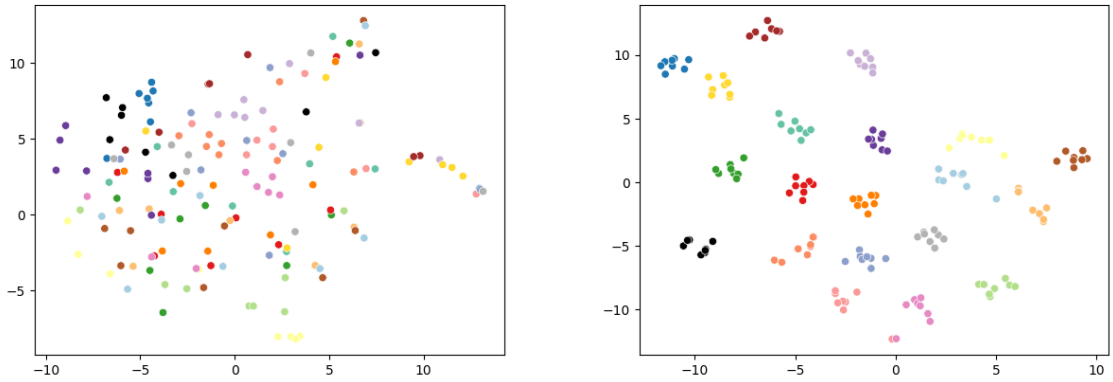
We conduct ablation studies to investigate the impact of each objective in our framework.

In this experiment, we remove each training objective respectively from our original model and get four pre-trained multilingual language models. Compared to the original MMKD, we can measure the performance gains of each training objective.

We report the average results across languages we evaluated on benchmark datasets in Table 5. We employ identical training setups to minimize the effect of other factors. We observe performance drops on ablated models across all the evaluation benchmarks.

On XNLI benchmark, MMKD outperforms other ablated models by 1.3%-2.0%. This result demonstrates that removing either proposed alignment objective will lead to less semantic knowledge. We can observe that performance drops dramatically on PAWS-X benchmark without the structure-level training objective. This indicates that aligning knowledge correlation between teacher and student models can benefit obtaining semantic information to distinguish sentences with similar words.

Similar to findings for XNLI and PAWS-X, F1 score on XQuAD benchmark worsens by 0.4% to



(a) representations from mBERT

(b) representations from MMKD

Figure 2: tSNE plots of 20 sentence representations on XNLI 15-way parallel corpus. We take eight languages used in our pre-training procedure. Dots in the same color are the representation of English and representation of its translations into other seven languages.

Models	XNLI	PAWS-X	XQuAD
Metrics	Acc	Acc	F1
MMKD	**75.4**	**88.6**	**70.1**
-XWCL	74.1	86.4	69.7
-TLM	73.5	87.5	68.5
-SentA	73.4	86.4	61.2
-StrucA	73.9	85.7	69.2

Table 5: Zero-shot cross-lingual results on evaluation benchmarks. MMKD indicates that the original model was pre-trained with all proposed objectives. - indicates the training objective which is removed from our framework, and thus the model is pre-trained using the other three training objectives.

8.9% without each training objective. Sentence-level alignment has a great impact on this question answering task.

In conclusion, each proposed training objective provides various semantic and structure knowledge and contributes to performance improvement.

4.5 Visualization of Representations

To further assess the effectiveness of our proposed alignment method, we visualize the sentence representations of MMKD and original mBERT using t-SNE (Van der Maaten and Hinton, 2008). We utilize a 15-way corpus provided by XNLI (Conneau et al., 2018). This corpus contains 10,000 sentences and their translations in fifteen languages. We randomly select 20 sentences and their translations in seven languages used in the pre-training procedure.

Figure 2 shows t-SNE plots of these sentence representations. Each dot represents a sentence representation produced by the multilingual language model. We treat [CLS] token embedding of last hidden states as the sentence representation. The dots in the same color are translations from the same sentence; thus, we have 8 dots in each color. Figure 2a and Figure 2b show representation t-SNE projections from mBERT and MMKD respectively. We observe that semantically similar sentences from different languages are clustered in vector space by MMKD, while these representations from mBERT do not follow this trend.

This visualization result confirms that our alignment method makes semantically similar sentences closed in the vector space even though they are from different languages. This result also proves the effectiveness of our method for transferring semantic knowledge from English to other languages.

5 Conclusion

In this work, we propose a Multi-level Multilingual Knowledge Distillation method to pre-train the multilingual language model - mBERT. We propose four training objectives to align token-, word-, and sentence-level information from parallel corpora, and we also learn knowledge correlation between teacher and student models. Compared to existing studies, we require fewer computational resources and less training time. In the zero-shot cross-lingual setting, our model outperforms models of similar size and reduces the performance gap to larger models across XNLI, PAWS-X, and XQuAD benchmarks. Experimental results show that MMKD obtains significant performance gains on low-resource languages and does well on Latin-

based languages. Visualization result shows that semantic relationships among sentences have been successfully transferred from English to other languages. Future work could extend our approach to other larger multilingual language models.

Limitations

In order to adopt rich vector space properties, we utilize English BERT as our teacher model during the pre-training procedure. MMKD achieves impressive results on Indo-European languages that are closed to English, while performance on languages from other language families that are distantly related to English get less improved. For example, Arabic and Chinese are high-resource languages whose performance across three evaluation tasks is lower than those of Indo-European languages. However, they still have performance gains compared to models of same size. Future work could consider combining multiple teacher models covering various language families.

Ethics Statement

The authors of this work follow the ACL Code of Ethics. This work complies with the ACL Ethics Policy.

Acknowledgements

The work was supported in part by the U.S. National Science Foundation (NSF) under Grant MRI-2018069 and Grant SES-2031002 to Feng Luo, Long Cheng, and Hongxin Hu.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

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*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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*.

Fei Ding, Yin Yang, Hongxin Hu, Venkat Krovi, and Feng Luo. 2020. Multi-level knowledge distillation via knowledge alignment and correlation. *arXiv preprint arXiv:2012.00573*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2020a. Explicit alignment objectives for multilingual bidirectional encoders. *arXiv preprint arXiv:2010.07972*.

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

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

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*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8854–8861.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2021. Tacl: Improving bert pre-training with token-aware contrastive learning. *arXiv preprint arXiv:2111.04198*.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020. Structure-level knowledge distillation for multilingual sequence labeling. *arXiv preprint arXiv:2004.03846*.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. *arXiv preprint arXiv:2007.15960*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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*.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9386–9393.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.