Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining

Chih-chan Tien *
University of Chicago
cctien@uchicago.edu

Shane Steinert-Threlkeld
University of Washington
shanest@uw.edu

Abstract

This work presents methods for learning cross-lingual sentence representations using paired or unpaired bilingual texts. We hypothesize that the cross-lingual alignment strategy is transferable, and therefore a model trained to align only two languages can encode multilingually more aligned representations. We thus introduce *dual-pivot transfer*: training on one language pair and evaluating on other pairs. To study this theory, we design unsupervised models trained on unpaired sentences and single-pair supervised models trained on bitexts, both based on the unsupervised language model XLM-R with its parameters frozen. The experiments evaluate the models as universal sentence encoders on the task of unsupervised bitext mining on two datasets, where the unsupervised model reaches the state of the art of unsupervised retrieval, and the alternative single-pair supervised model approaches the performance of multilingually supervised models. The results suggest that bilingual training techniques as proposed can be applied to get sentence representations with multilingual alignment.

1 Introduction

Cross-lingual alignment as evaluated by retrieval tasks has been shown to be present in the representations of recent massive multilingual models which are not trained on bitexts (Pires et al., 2019; Conneau et al., 2020b). Other studies further show that sentence representations with higher cross-lingual comparability can be achieved by training a cross-lingual mapping (Aldarmaki and Diab, 2019) or fine-tuning (Cao et al., 2020) for every pair of languages. These two lines of research show that, on the one hand, multilingual alignment arises from training using monolingual corpora alone, and, on the other, bilingual alignment can be enhanced by training on bitexts of specific language pairs.

Combining these insights yields a question: can training with bilingual corpora help improve *multilingual* alignment? Given a language model encoding texts in different languages with some shared structure already, we can expect that the model further trained to align a pair of languages will take advantage of the shared structure and will therefore generalize the alignment strategy to other language pairs. From a practical point of view, bitexts for some pairs of languages are more abundant than others, and it is therefore efficient to leverage data from resource-rich pairs for the alignment of resource-poor pairs in training multilingual language models.

To better understand the cross-lingual structure from the unsupervised models, we also ask the following question: how can multilingual alignment information be extracted from the unsupervised language models? Unsupervised multilingual models out-of-the-box as sentence encoders fall short of their supervised counterparts such as LASER (Artetxe and Schwenk, 2019b) in the task of bitext mining (Hu et al., 2020). The discovery of cross-lingual structure in the hidden states in the unsupervised model (Pires et al., 2019; Conneau et al., 2020b), however, raises the possibility that with relatively light post-training for better extraction of deep features, the unsupervised models can generate much more multilingually aligned representations.

In this paper, we address both questions with the design of *dual-pivot transfer*, where a model is trained for bilingual alignment but tested for multilingual alignment. And we hypothesize that training to encourage similarity between sentence representations from two languages, the dual pivots, can help generate more aligned representations not only for the pivot pair, but also for other pairs.

In particular, we design and study a simple extraction module on top of the pretrained multilingual language model XLM-R (Conneau et al.,

---

*Work done while at the Department of Linguistics at the University of Washington.

2020a). To harness different training signals, we propose two training architectures. In the case of training on unpaired sentences, the model is encouraged by adversarial training to encode sentences from the two languages with similar distributions. In the other case where bitexts of the two pivot languages are used, the model is encouraged to encode encountered parallel sentences similarly. Both models are then transferred to language pairs other than the dual pivots. This enables our model to be used for unsupervised bitext mining, or bitext mining where the model is trained only on parallel sentences from a single language pair.

The experiments show that both training strategies are effective, where the unsupervised model reaches the state of the art on completely unsupervised bitext mining, and the one-pair supervised model approaching the state-of-the-art multilingually-supervised language models in one bitext mining task.

Our contributions are fourfold:

- This study proposes effective methods of bilingual training using paired or unpaired sentences for sentence representation with multilingual alignment. The strategies can be incorporated in language model training for greater efficiency in the future.

- The work demonstrates that the alignment information in unsupervised multilingual language models is extractable by simple bilingual training of a light extraction module (without fine-tuning) with performance comparable to fully supervised models and reaching the state of the art of unsupervised models.

- The models are tested using a new experimental design—*dual-pivot transfer*—to evaluate the generalizability of a bilingually-supervised sentence encoder to the task of text mining for other language pairs on which it is not trained.

- This study shows that unsupervised bitext mining has strong performance which is comparable to bitext mining by a fully supervised model, so the proposed techniques can be applied to augment bilingual corpora for data-scarce language pairs in the future.

2 Related work

Alignment with adversarial nets This work follows the line of previous studies which use adversarial networks (GANs) (Goodfellow et al., 2014) to align cross-domain distributions of embeddings without supervision of paired samples, in some cases in tandem with cycle consistency (Zhu et al., 2017), which encourages representations "translated" to another language then "translated" back to be similar to the starting representations. Conneau et al. (2018)'s MUSE project trains a linear map from the word-embedding space of one language to that of another using GANs, the method of which is later applied to an unsupervised machine translation model (Lample et al., 2018a). Cycle consistency in complement to adversarial training has been shown to be effective in helping to learn cross-lingual lexicon induction (Zhang et al., 2017; Xu et al., 2018; Mohiuddin and Joty, 2020). Our work is the first to our knowledge to apply such strategy of adversarial training and cycle consistency to the task of bitext mining.

Alignment with pretrained LMs We adopt the training strategy aforementioned on top of pretrained multilingual language models, the extractability of multilingual information from which has been studied in several ways. Pires et al. (2019) find multilingual alignment in the multilingual BERT (mBERT) model (Devlin et al., 2019) pretrained on monolingual corpora only, while Conneau et al. (2020b) identify shared multilingual structure in monolingual BERT models. Other work studies the pretrained models dynamically by either fine-tuning the pretrained model for cross-lingual alignment (Cao et al., 2020) or learning cross-lingual transformation (Aldarmaki and Diab, 2019) with supervision from aligned texts. Recently, Yang et al. (2020) use multitask training to train multilingual encoders focusing on the performance on retrieval, and Reimers and Gurevych (2020) use bitexts to tune multilingual language models and to distill knowledge from a teacher model which has been tuned on paraphrase pairs. Also, Chi et al. (2021a) pretrain an alternative XLM-R on a cross-lingual contrastive objective. Our work falls in the line of exploring multilinguality of pretrained models with a distinct emphasis on investigating the multilingual structure induced by bilingual training without fine-tuning or alternative pretraining.
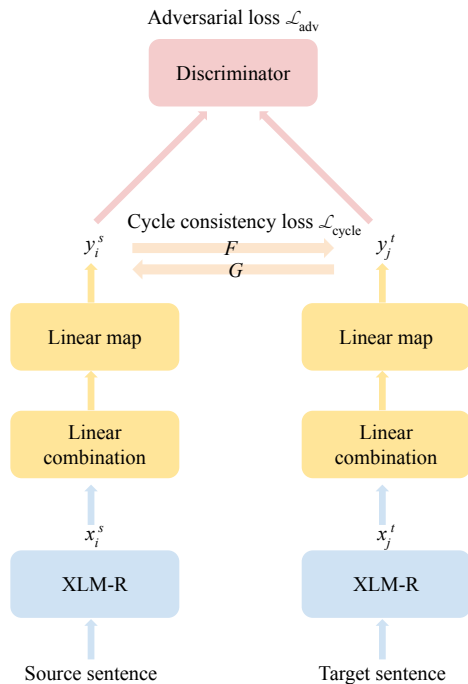
Figure 1: Schematic depiction of the unsupervised model with adversarial and cycle consistency losses.

Unsupervised parallel sentence mining The evaluation task of our work is bitext mining without supervision from any bitexts or from bitexts of the pair of languages of the mining task. Such experiments have been explored previously. Hangya et al. (2018) show that unsupervised bilingual word embeddings are effective on bitext mining, and Hangya and Fraser (2019) further improve the system with a word-alignment algorithm. Kiros (2020) trains a lensing module over mBERT for the task of natural language inference (NLI) and transfers the model to bitext mining. Keung et al. (2020)'s system uses bootstrapped bitexts to fine-tune mBERT, while Kvapilíková et al. (2020)'s system uses synthetic bitexts from an unsupervised machine translation system to fine-tune XLM (Conneau and Lample, 2019). Results from the three aforementioned studies are included in Section 5 for comparisons. Methodologically, our approach differs from the above in that our system is based on another pretrained model XLM-R (Conneau et al., 2020a) without fine-tuning, for one of the goals of the study is to understand the extractability of the alignment information from the pretrained model; and our model receives training signals from existing monolingual corpora or bitexts, instead of from NLI, bootstrapped, or synthesized data.

3 Model

3.1 A linear combination and a linear map

The model as an encoder generates fixed-length vectors as sentence representations from the the hidden states of the pretrained multilingual language model XLM-R (Conneau et al., 2020a). Formally, given a sentence $\pi_{i,\gamma}$ in language $\gamma \in \{s, t\}$, with the pretrained language model producing features x_i^γ of l layers, sequence length q, and embedding size d, the extraction module $f(\cdot)$ generates a sentence embedding y_i^γ of fixed size d based on the features x_i^γ, or

$$f(x_i^\gamma) = y_i^\gamma, \quad x_i^\gamma \in \mathbb{R}^{l \times q \times d} \text{ and } y_i^\gamma \in \mathbb{R}^d.$$

With the parameters of XLM-R frozen, within the extraction module $f(\cdot)$ are only two trainable components. The first is an ELMo-style trainable softmax-normalized weighted linear combination module (Peters et al., 2018), and the second being a trainable linear map. The linear combination module learns to weight the hidden states of every layer l of the pretrained language model and output a weighted average, on which a sum-pooling layer is then applied to q embeddings. And then the linear map takes this bag-of-word representation and produces the final sentence representation y_i^γ of the model.

3.2 Adversarial learning with unpaired texts

Monolingual corpora in different languages with language labels provide the signal for alignment if the semantic contents of the utterances share similar distributions across corpora. In order to exploit this information, we introduce to the model adversarial networks (Goodfellow et al., 2014) with cycle consistency (Zhu et al., 2017), to promote similarity in the distribution of representations.

As is usual in GANs, there is a discriminator module $d(\cdot)$, which in this model consumes the representation y_i^γ and outputs continuous scores for the language identity γ of the sentence $\pi_{i,\gamma}$. Following Romanov et al. (2019), as inspired by Wasserstein-GAN (Arjovsky et al., 2017), the loss of the discriminator $\mathcal{L}_{\text{disc}}$ is the difference between the unnormalized scores instead of the usual cross-entropy loss, or

$$\mathcal{L}_{\text{disc}} = d(y_i^s) - d(y_j^t).$$

And the adversarial loss $\mathcal{L}_{\text{adv}} = -\mathcal{L}_{\text{disc}}$ updates the parameters of the extraction module $f(\cdot)$ to

encourage it to generate encodings abstract from language-specific information.

Adversarial training helps learning aligned encodings across languages at the distributional level. At the individual level, however, the model is not constrained to generate encodings which are both aligned and discriminative (Zhu et al., 2017). In particular, a degenerate encoder can produce pure noise which is distributively identical across languages. A cycle consistency module inspired by Zhu et al. (2017) is therefore used to constrain the model to encode with individual-discerning alignment. Cycle consistency is also reminiscent of the technique of using back-translation for unsupervised translation systems (Lample et al., 2018b).

In this model, a trainable linear map $F(\cdot)$ maps elements from the encoding space of one language to the space of the other, and another linear map $G(\cdot)$ operates in the reverse direction. The cycle loss so defined is used to update parameters for both of the cycle mappings and the encoder:

$$\mathcal{L}_{\text{cycle}} = h(y_i^s, G(F(y_i^s))) + h(F(G(y_j^t)), y_j^t)$$

where h is the triplet ranking loss function which sums the hinge costs in both directions:

$$h(a,b) = \sum_n \max(0, \alpha - \text{sim}(a,b) + \text{sim}(a_n, b))$$
$$+ \max(0, \alpha - \text{sim}(a,b) + \text{sim}(a, b_n)),$$

where the margin α and the number of negative samples n are hyperparameters, and $\text{sim}(\cdot)$ is cosine similarity. The loss function h encourages the model to encode similar representations between positive pairs (a, b) and dissimilar representations between negative pairs (a_n, b) and (a, b_n), where a_n and b_n are sampled from the embeddings in the mini-batch. Based on the findings that the hard negatives, or non-translation pairs of high similarity between them, are more effective than the sum of negatives in the ranking loss (Faghri et al., 2018), our system always includes in the summands the costs from the hardest negatives in the mini-batch along with the costs from any other randomly sampled ones.

The full loss of the unsupervised model is

$$\mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{cycle}},$$

with a hyperparameter λ. This unsupervised model is presented schematically with Figure 1.

Hyperparameter		values
output dimension	d	$\{1024\}$
# negative samples	n	$\{1, 2, 4\}$
margin value	α	$\{0, 0.2, 0.4\}$
weight for cycle loss	λ	$\{1, 5, 10\}$
discriminator step times	κ	$\{1, 2\}$

Table 1: Hyperparameters and experimented values.

3.3 Learning alignment with bitexts

In addition to the completely unsupervised model, we also experiment with a model which is supervised with bitext from one pair of languages and then transferred to other pairs. In this set-up, instead of using cyclical mappings, bitexts provide the alignment signal through the ranking loss directly, so the loss for the supervised model is

$$\mathcal{L}_{\text{sup}} = h(y_i^s, y_i^t),$$

where y_i^s and y_i^t are representations of parallel sentences.

4 Training

The model is trained with the Adam optimizer (Kingma and Ba, 2014) and learning rate 0.001 with the parameters in XLM-R frozen. Our training program is built upon AllenNLP (Gardner et al., 2018), HuggingFace Transformers (Wolf et al., 2020), and PyTorch (Paszke et al., 2019). The code for this study is released publicly.[1]

For adversarial training, the discriminator is updated κ times for every step of backpropagation to the encoder. Other hyperparameters include the dimension of the output representations d, number of negative samples n, margin value α, and weight of the cycle loss λ. The hyperparameters and the the values which are experimented with are summarized in Table 1. We empirically determine the hyperparameters among experimented values, and report their values in specific evaluation sections.

The bilingual corpora used to train the encoder is taken from OPUS (Tiedemann, 2012) as produced for the training of XLM (Conneau and Lample, 2019).[2] We experimented with two language pairs for training the model—Arabic-English (ar-en) and

---

[1] The repository at https://github.com/cctien/bimultialign.

[2] We use the script https://github.com/facebookresearch/XLM/blob/main/get-data-para.sh to get the corpora MultiUN (Eisele and Chen, 2010) and EUbookshop, where each training corpus we use is of 9 million sentences.

Model	F1 score (%)				xx↔en
	Average	de	fr	ru	zh
Unsupervised					
Model (ar-en unsup.)	81.8	84.1	77.9	**87.9**	**77.1**
Model (de-en unsup.)	**82.4**	**91.4**	75.6	86.1	76.5
XLM-R L16-boe	68.7	75.4	65.0	75.6	59.0
Kvapilíková et al. (2020)	75.8	80.1	**78.8**	77.2	67.0
Keung et al. (2020)	69.5	74.9	73.0	69.9	60.1
Kiros (2020)	51.7	59.0	59.5	47.1	41.1
One-pair supervised					
Model (ar-en bitexts sup.)	89.1	91.7	89.2	90.1	85.6
Model (de-en bitexts sup.)	**89.6**	**92.5**	**89.6**	**90.3**	**85.8**
Fully Supervised					
LASER	92.8	95.4	92.4	92.3	91.2
LaBSE	**93.5**	**95.9**	**92.5**	**92.4**	**93.0**
XLM-R+SBERT	88.6	90.8	87.1	88.6	87.8

Table 2: F1 scores on the BUCC bitext mining text. Simple average of scores from 4 tasks reported in the second column. Highest scores in their groups are bolded.

German-English (de-en)—to explore potential effects of the choice of the dual-pivots. After being trained, the encoder is evaluated on two tasks of bitext mining between texts in English and in another language. Additionally, we train the models with the pivot pair of Arabic-German (ar-de), which does not include English, to be evaluated on the second task.

5 Evaluations

Four models, two unsupervised and two one-pair supervised trained on either of the two language pairs, are evaluated on two bitext mining or retrieval tasks of the BUCC corpus (Zweigenbaum et al., 2018) and of the Tatoeba corpus (Artetxe and Schwenk, 2019a).

5.1 Baselines and comparisons

Unsupervised baselines The XLM-R (Conneau et al., 2020a) bag-of-embedding (boe) representations out-of-the-box serve as the unsupervised baseline. We identify the best-performing among the layers of orders of multiples of 4, or layer $L \in \{0, 4, 8, 12, 16, 20, 24\}$, as the baseline. In the case of BUCC mining task, for example, the best-performing baseline model is of layer 16 and denoted by XLM-R L16-boe.

Results from Kiros (2020), Keung et al. (2020), and Kvapilíková et al. (2020), as state-of-the-art models for unsupervised bitext mining from pretrained language models, are included for compari-

son (see Section 2 for a description of them).

Fully supervised models LASER (Artetxe and Schwenk, 2019b) and LaBSE (Feng et al., 2020), both fully supervised with multilingual bitexts, are included for comparisons. LASER is an LSTM-based encoder and translation model trained on parallel corpora of 93 languages, and is the earlier leading system on the two mining tasks. LaBSE on the other hand is a transformer-based multilingual sentence encoder supervised with parallel sentences from 109 languages using the additive margin softmax (Wang et al., 2018) for the translation language modeling objective, and has state-of-the-art performance on the two mining tasks. Finally, XLM-R+SBERT from Reimers and Gurevych, 2020 is XLM-R fine-tuned to align representations of bitexts of 50 language pairs and to distill knowledge from SBERT (Reimers and Gurevych, 2019) fine-tuned on English paraphrase pairs.

5.2 BUCC

The BUCC corpora (Zweigenbaum et al., 2018), consist of 95k to 460k sentences in each of 4 languages—German, French, Russian, and Mandarin Chinese—with around 3% of such sentences being English-aligned. The task is to mine for the translation pairs.

Margin-based retrieval The retrieval is based on the margin-based similarity scores (Artetxe and Schwenk, 2019a) related to CSLS (Conneau et al., 2018),

$$\text{score}(y^s, y^t) = \text{margin}(\text{sim}(y^s, y^t), \text{scale}(y^s, y^t))$$

$$\text{scale}(y^s, y^t) =$$

$$\sum_{z \in \text{NN}_k(y^s)} \frac{\text{sim}(y^s, z)}{2k} + \sum_{z \in \text{NN}_k(y^t)} \frac{\text{sim}(y^t, z)}{2k},$$

where $\text{NN}_k(y)$ denotes the k nearest neighbors of y in the other language. Here we use $k = 4$ and the ratio margin function, or $\text{margin}(a, b) = a/b$, following the literature (Artetxe and Schwenk, 2019b). By scaling up the similarity associated with more isolated embeddings, margin-based retrieval helps alleviate the hubness problem (Radovanovic et al., 2010), where some embeddings or hubs are nearest neighbors of many other embeddings with high probability.

Following Hu et al. (2020), our model is evaluated on the training split of the BUCC corpora, and the threshold of the similarity score cutting

off translations from non-translations is optimized for each language pair. While Kvapilíková et al. (2020) and Kiros (2020) optimize for the language-specific mining thresholds as we do here, Keung et al. (2020) use a prior probability to infer the thresholds. And different from all other baselines or models for comparisons presented here, Kvapilíková et al. (2020)'s model is evaluated upon the undisclosed test split of the BUCC corpus.

Results F1 scores on the BUCC dataset presented in Table 2 demonstrate that bilingual alignment learned by the model is transferable to other pairs of languages. The hyperparameter values of the unsupervised model presented in the table are $n = 1, \alpha = 0, \lambda = 5, \kappa = 2$, and those of the supervised model are $n = 1, \alpha = 0$.

The adversarially-trained unsupervised model outperforms the unsupervised baselines and nearing the state of the art, and is thus effective in extracting sentence representations which are sharable across languages. The choice of pivot pairs shows effects on the unsupervised models, with the model trained on the de-en texts performing better than that on the ar-en texts at mining for parallel sentences between English and German by 7 points. The results suggest that while alignment is transferable, the unsupervised model can be further improved for multilingual alignment by being trained on multilingual texts of more than two pivots.

The one-pair supervised model trained with bitexts of one pair of languages, on the other hand, performs within a 6-point range of the fully supervised systems, which shows that much alignment information from unsupervised pretrained models is recoverable by the simple extraction module. Noticeably, the model supervised with ar-en bitexts but not from the four pairs of the task sees a 20-point increase from the plain XLM-R, and the choice of dual pivots does not have significant effects on the supervised model.

5.3 Tatoeba

We also measure the parallel sentence matching accuracy over the Tatoeba dataset (Artetxe and Schwenk, 2019b). This dataset consists of 100 to 1,000 English-aligned sentence pairs for 112 languages, and the task is to retrieve the translation in one of the target languages given a sentence in English using absolute similarity scores without margin-scaling.

Model	Average accuracy (%)		
	36	41	112
Unsupervised			
Model (ar-en unsup.)	73.4	70.8	55.3
Model (ar-de unsup.)	71.5	68.9	53.3
Model (de-en unsup.)	**74.2**	**72.0**	**56.0**
XLM-R L12-boe	54.3	53.1	39.6
Kvapilíková et al. (2020)	——	50.2	——
One-pair supervised			
Model (ar-en bitexts sup.)	79.6	78.3	61.1
Model (ar-de bitexts sup.)	74.1	71.8	56.0
Model (de-en bitexts sup.)	**80.4**	**78.9**	**62.4**
Supervised with 50+ pairs			
LASER	85.4	79.1	66.9
LaBSE	**95.0**	——	**83.7**
XLM-R+SBERT	86.2	84.3	67.1

Table 3: Average accuracy scores on the Tatoeba dataset in three average groups. Highest scores in their groups are bolded.

Results Matching accuracy for the retrieval task of the Tatoeba dataset are presented in Table 3. Following Feng et al., 2020, average scores from different groups are presented to compare different models. The 36 languages are those selected by Xtreme (Hu et al., 2020), and the 41 languages are those for which results are presented in Kvapilíková et al., 2020. The hyperparameter values of the unsupervised model presented in the table are $n = 2, \alpha = 0.2, \lambda = 10, \kappa = 2$, and those of the supervised model are $n = 1, \alpha = 0$.

The unsupervised model outperforms the baselines by roughly 20 points, and the one-pair supervised model performs close to the the supervised model LASER but falls shorts by around 10 to 20 points to the other supervised model LaBSE. When one of the two pivot languages is English, the choice of the pivot does not show much difference on this task on average. While the models trained on ar-de (where neither pivot language is English) still exhibits strong transfer performance, there is a drop of around 2 to 6 points from the models where English is one of the pivot languages (ar-en and de-en).

6 Analysis

To understand the factors affecting the performance of the model, we consider several variants. All models presented in this section are trained with the same hyperparameters presented in the evaluation section above using either de-en corpora or corpora of multiple language pairs (Section 6.3).

Model	Tatoeba 36
Unsupervised	
XLM-R average-boe	54.9
XLM-R L12-boe	54.3
$\mathcal{L} = \mathcal{L}_{\text{adv}}$	
linear combination	34.7
linear map	0.1
linear combination + linear map	2.1
$\mathcal{L} = \mathcal{L}_{\text{cycle}}$	
linear combination	55.4
linear map	69.8
linear combination + linear map	68.0
$\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{cycle}}$	
linear combination	47.4
linear map	70.5
linear combination + linear map	74.2
One-pair supervised	
$\mathcal{L} = \mathcal{L}_{\text{sup}}$	
linear combination	67.5
linear map	80.1
linear combination + linear map	80.4

Table 4: Ablation results of the accuracy (%) on the Tatoeba averaged across 36 languages.

6.1 Ablation

We ablate from the model the trainable components—the weighted linear combination and the linear map—as well as the two training losses, $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{cycle}}$, of the unsupervised model. When the weighted linear combination is ablated, we use the unweighted average of embeddings across layers.

We evaluate on the average accuracy over the 36 languages of the Tatoeba corpus. The results in Table 4 show a few interesting trends. First, the cycle consistency loss is essential for the unsupervised model, as can be seen by the very low performance when only $\mathcal{L}_{\text{adv}}$ is used. Secondly, the linear map plays a larger role than the linear combination in extracting alignment information in the unsupervised model: in both conditions with cycle consistency loss, the linear map alone outperforms the linear combination alone, and in the condition with only cycle consistency loss, the linear map alone does best. Finally, in the one-pair supervised model, the linear combination module alone shows gains of 13 points from the baseline but does not produce gains when trained along with a linear map.
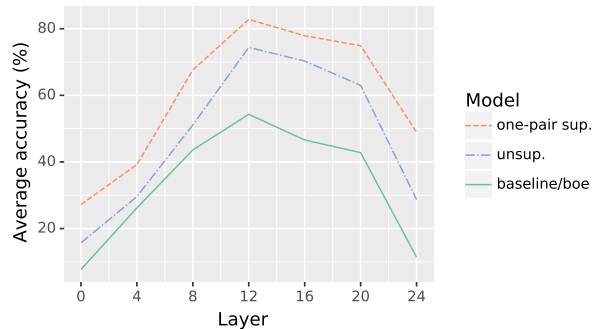

Figure 2: Average accuracies of 36 languages of the Tatoeba task where the model is trained with representations from XLM-R layers of orders which are multiples of 4.

Model	Model with # of pairs				
	1	2	4	8	16
Unsupervised	74.2	74.9	75.2	77.5	77.1
Supervised	80.4	80.0	80.0	83.3	85.7

Table 5: Models with multiple pairs with accuracy (%) on Tatoeba averaged across 36 languages.

6.2 Single-layer representations

Previous studies show that representations from different layers differ in their cross-lingual alignment (Pires et al., 2019; Conneau et al., 2020b). To understand this phenomenon in the present setting, we take layers whose orders are multiples of 4 and train the model with representations from one single layer without combining embeddings from different layers.

The average accuracy scores over 36 languages in Tatoeba summarized in Figure 2 show that the alignment information is most extractable deep in the middle layers, corroborating the findings from the previous work (Kvapilíková et al., 2020; Litschko et al., 2021; Chi et al., 2021b). The model trained with the best-performing layer shows similar or higher scores than the full model with learned linear combination, which is consistent with the findings from the ablation that the learned linear combination is not essential for extracting alignment information.

6.3 Training with multiple language pairs

It is possible that a model trained on texts from more pairs of languages may improve upon the multilingual alignment so far demonstrated. To test this, we trained the model with the same hyper-

Model	F1 score (%) xx↔en			
	de	fr	ru	zh
Unsupervised				
Optimized thresholds	91.4	75.6	86.1	76.5
Dual-pivot threshold	91.4	73.5	84.9	75.8
One-pair supervised				
Optimized thresholds	92.5	89.6	90.3	85.8
Dual-pivot threshold	92.5	89.6	90.2	85.2

Table 6: F1 scores on the BUCC training split with the model trained with de-en texts. The system with optimized thresholds tunes the threshold for each pair, as in Table 2; the dual-pivot system uses the threshold from the de-en pair for all four pairs.

Model	ar-ar	en-en	es-es	en-ar	en-es	en-tr
Unsup.						
XLM-R	47.2	58.5	53.0	31.2	17.3	28.1
Model	51.7	74.5	63.8	42.0	41.9	37.9
One-pair sup.						
Model	54.7	67.5	65.0	43.7	39.8	44.0
Other systems						
LASER	68.9	77.6	79.7	66.5	57.9	72.0
LaBSE	69.1	79.4	80.8	74.5	65.5	72.0
SBERT	79.6	88.8	86.3	82.3	83.1	80.9

Table 7: Spearman correlation (%) on multilingual STS 2017 (Cer et al., 2017) of the unsupervised and supervised models trained on the de-en pair.

parameters as in the previous section but on texts from multiple languages,[3] where each multi-pair model is trained on 16 million total sentences. The results are in Table 5. There is an aggregate 3 point improvement for the unsupervised model and 5 point for the supervised model. The results suggest that with our models, one bilingual pivot is capable of extracting much transferable multilingual representations from XLM-R, but using more pivots can still improve the transferability of the representations to some extent.

6.4 Threshold transfer

Previous work observes that in the BUCC mining task, the thresholds optimized for different language pairs are close to one another, suggesting that one can tune the threshold on high-resource pairs and use the system to mine other language pairs (Kiros, 2020). We examine the mining performance on the BUCC dataset of two threshold schemes: optimized thresholds, where thresholds are optimized for each language pair, and the dual-pivot threshold, where the threshold optimized for the pivot pair, in this case German-English, is used to mine all languages.

The scores from these two schemes on BUCC are summarized in Table 6. The results show that the thresholds optimized for the pivot pair transfer to other pairs with at most a 2-point decrease in the F1 score, and that the results of the two schemes are almost identical to the one-pair supervised model. These experiments corroborate the

previous observation and demonstrate yet another case for leveraging texts from resource-rich pairs for unsupervised mining of other language pairs.

6.5 Multilingual semantic textual similarity

To test whether our method of training language-agnostic sentence encoders encourages meaning-based representations, we evaluate the models on the multilingual semantic textual similarity (STS) 2017 (Cer et al., 2017) with Spearman correlation reported in Table 7. All evaluation pairs on average see about 10 percentage-point increases from baseline (XLM-R L12-boe) for both models. Yet the gaps between our models and the fully supervised systems suggest that supervision with more language pairs and more trainable parameters likely encourages sentence representations to be closer to what humans see as meaning.

7 Conclusion

This work shows that training for bilingual alignment benefits multilingual alignment for unsupervised bitext mining. The unsupervised model shows the effectiveness of adversarial training with cycle consistency for building multilingual language models, and reaches the state of the art of unsupervised bitext mining. Both unsupervised and one-pair supervised models show that significant multilingual alignment in an unsupervised language model can be recovered by a linear mapping, and that combining monolingual and bilingual training data can be a data-efficient method for promoting multilingual alignment. Future work may combine both the supervised and the unsupervised techniques to attain sentence embeddings with stronger multilingual alignment through the transferability of bilingual alignment demonstrated

---

[3]The 16 pairs are those between `en` and `ar`, `bg`, `de`, `el`, `es`, `fa`, `fr`, `hi`, `id`, `ru`, `sw`, `th`, `tr`, `ur`, `vi`, `zh`.

in this work, and such work will benefit tasks involving languages of low resources in bitexts.

References

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. 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)*, pages 3906–3911, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of Machine Learning Research*, volume 70, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021b. Xlm-e: Cross-lingual language model pre-training via electra. *arXiv preprint arXiv:2106.16138*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger,

editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *International Workshop on Spoken Language Translation*.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2020. Unsupervised bitext mining and translation via self-trained contextual embeddings.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Jamie Kiros. 2020. Contextual lensing of universal sentence representations.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Proceedings of ECIR*.

Tasnim Mohiuddin and Shafiq Joty. 2020. Unsupervised word translation with adversarial autoencoder. *Computational Linguistics*, 46(2):257–288.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Alexey Romanov, Anna Rumshisky, Anna Rogers, and David Donahue. 2019. Adversarial decomposition of text representation. 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)*, pages 815–825, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul,

Turkey. European Language Resources Association (ELRA).

F. Wang, J. Cheng, W. Liu, and H. Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. A multilingual dataset for evaluating parallel sentence extraction from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).