Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining

Chih-chan Tien

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2020

Committee:

Shane Steinert-Threlkeld

Gina-Anne Levow

Program Authorized to Offer Degree:
Computational Linguistics

University of Washington

Abstract

Bilingual alignment transfers to multilingual alignment
for unsupervised parallel text mining

Chih-chan Tien

Chair of the Supervisory Committee:
Doctor Shane Steinert-Threlkeld
Department of Linguistics

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. The method of transferring bilingual alignment between two pivot languages to multilingual alignment among other languages is novel and we call this method *dual-pivot transfer*. To study the applicability of the transfer, we train an unsupervised model with unpaired sentences and another single-pair supervised model with bitexts, both based on the unsupervised language model XLM-R. 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 higher multilingual alignment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am greatly in debt to my advisor Doctor Shane Steinert-Threlkeld for his kind and enlightening guidance, without which most progress would have been impossible. I am extremely grateful to Professor Gina-Anne Levow for her substantial advice which makes the work much better. I thank Profesor Luke Zettlemoyer for being inspiring and for giving sound suggestions.

DEDICATION

To Tai-jun, Han-mei, and Chih-chin.

Chapter 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 with 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 with 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 (Joshi et al., 2020), 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 easily can multilingual alignment information be extracted from the unsupervised language models? Unsupervised multilingual models used 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 a pretrained unsupervised language model XLM-R (Conneau et al., 2020a). To harness different training signals, we propose two training architectures. In the case of training with 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 with high similarity. Both models are then transferred to language pairs other than the dual pivots. This enables the model to be used for unsupervised bitext mining, or bitext mining with the model trained only with parallel sentences from a different language pair.

The model as a sentence encoder is evaluated on the task of bitext mining on two datasets, the BUCC dataset (Zweigenbaum et al., 2018) and the Tatoeba dataset (Artetxe and Schwenk, 2019b). For each language pair in the dataset, the model computes the similarity scores for all sentence pairs in the corpus, and then the pairs are ranked based on the scores to determine whether or not a particular pair is a translation. Table 1.1 shows two example pairs and their similarity scores.

English sentence	German sentence	score	parallel?
How many hours of sleep do you need?	Wie viele Stunden Schlaf brauchst du?	0.81	Y
This is a pun.	Das ist eine Sackgasse.	0.26	N

Table 1.1: Examples sentences and the output similarity scores. The first pair is a parallel sentence, wheres the second is not. All four example sentences are from the Tatoeba dataset (Artetxe and Schwenk, 2019b).

The experiments show that both training strategies are effective, with the unsupervised model reaching the state-of-the-art of completely unsupervised bitext mining, and the one-pair supervised

model approaching the state-of-the-art for multilingually-supervised language models in one mining task.

The contributions are fourfold:

- This study proposes effective methods of bilingual training using paired or unpaired sentences for sentence representation with higher 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 with performance comparable to the state-of-the-art fully supervised models.

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

- 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.

The thesis is structured with six chapters. Following the introduction in Chapter 1, we discuss the literature related to this work in Chapter 2. We describe the methodology for building and training the model in Chapter 3, the evaluation tasks and results in Chapter 4, and some analyses in Chapter 5. Chapter 6 finally concludes.

Chapter 2

RELATED WORK

The work of this thesis is at the intersection of several lines of work described in this chapter. Our model transforms sequence representations from pretrained neural language models (Section 2.1) using training signals from a contrastive loss function (Section 2.2) and an adversarial loss (Section 2.3). Through these methods, the work probes for the cross-lingual alignment structure of an unsupervised neural language models (Section 2.4), and studies the task of unsupervised bitext mining (Section 2.5).

2.1 Sequence representations from pretrained models

This work follows the established technique of using sequence representations from pretrained neural language models.

BERT (Devlin et al., 2019) prepends a special `[CLS]` token to the input sequence to a Transformer-based (Vaswani et al., 2017) model, and the embedding of the final hidden state of this special token is used as the sentence representation for sequence-level classification tasks during fine-tuning and evaluation. This approach has been used for other models based on the BERT architecture, including XLM-R (Conneau et al., 2020a).

Besides BERT's native CLS approach, some other sequence-to-vector strategies were explored. Reimers and Gurevych (2019) evaluate three pooling approaches, i.e., CLS, mean-pooling (bag-of-embeddings (BOE)), and max-pooling, for the sentence representations from the hidden states of BERT for the tasks of NLI and semantic textual similarity (STS). They show that while the choice of pooling strategies has little effects on the NLI task, both BOE and CLS significantly outperform the max-pooling method in the STS task.

ELMo (Peters et al., 2018) does not explicitly extract sentence representation from the deep

contextualizer, yet it has a trainable scalar mixture or linear combination module to weight and extract features from all layers. When applied on BERT, this feature-extraction approach, along with task-specific sentence-pair matching techniques, is on par with the fine-tuning approach on NLI tasks, but falls short on STS tasks (Peters et al., 2019).

Finally, Kiros (2020) studies task-specific extraction modules or *lenses*, which transform embeddings from the final layer of BERT into specializing embeddings suitable for different downstream tasks without fine-tuning the parameters of BERT. The architecture of the lenses are based on InferLite (Kiros and Chan, 2018). And depending on the use cases or contexts, different data and strategies are used to train the lenses. For instance, NLI benchmarks are used to train the lens for general-purpose monolingual sentence encodings, while parallel corpora are used to train the translation lens.

This study is an extension of Kiros (2020) with different objectives, which of this work are to train a single sentence encoder for both cross-lingual similarity and transferability of sentence encodings, and to explore the possibilities of dual-pivot transfer to languages with no available parallel sentences. My work also differs with theirs in the model design, for which I adopt the deep feature-extraction approach (Peters et al., 2018) to encourage the extraction module to learn all alignment information from every layer of the language model. To convert token-level representations to sequence-level representations, I use the BOE pooler, which was shown to be effective in previous work (Reimers and Gurevych, 2019).

2.2 Contrastive training for representation learning of texts

The triplet ranking loss with hinges and negative samples has been used to learn cross-domain alignment (Lazaridou et al., 2015). The goal of the loss is to encourage similarity between positive pairs while discouraging similarity between negative pairs when the negatives are hard enough or passing a threshold. And benefit of the threshold or the hinge is that it helps reduce the problem of hubness (Lazaridou et al., 2015), or the problems that some target embeddings or hubs are nearest neighbors of many other embeddings with high probability, which makes retrieving embeddings based on closet neighbors difficult and prone to errors. Below we discuss some use cases of hinge

losses and triplet ranking losses in tasks involving natural langauges.

In their seminal work, Collobert and Weston (2008) pioneer word embeddings trained with the hinge loss and use a single convolutional neural network which, given a sentence, outputs a variety of language processing predictions. The loss is for one sentence window of text s is formulated as

$$\mathcal{L} = \sum_{w} \max(0, 1 - f(s) + f(s^w)) \,,$$

where $f(\cdot)$ represents the neural network, and s^w is a sentence window where the middle word has been replaced by a negative sampled word w.

Recently, Conneau et al. (2018b) use the triplet ranking loss with negative sampling to align cross-lingual sentence embeddings derived from averaged word embeddings with the formula

$$\begin{aligned} \mathcal{L} = \sum_{n} & \max(0, \alpha + \mathrm{dist}(x, y) - \mathrm{dist}(x_n, y)) \\ & + \max(0, \alpha + \mathrm{dist}(x, y) - \mathrm{dist}(x, y_n)) \,, \end{aligned} \tag{2.1}$$

where n is the number of negative samples, α is the margin hyperparameter, dist is a distance measure, and x and y are sentence embeddings of the source and target language, respectively.

The common formulation of the triplet ranking loss is to sum over hinges from several negative examples. Yet Faghri et al. (2018) show that the ranking loss with the single hardest negative sample for each anchor performs better than that of sum of hinges of several negatives in the task of learning joint embeddings for images and caption texts. Formally, Faghri et al. (2018) suggest to use the max of hinges or

$$\begin{aligned} \mathcal{L} = & \max_{n} \max(0, \alpha + \mathrm{dist}(x, y) - \mathrm{dist}(x_n, y)) \\ & + \max_{n} \max(0, \alpha + \mathrm{dist}(x, y) - \mathrm{dist}(x, y_n)) \end{aligned} \tag{2.2}$$

in place of the sum of hinges or formula (2.1).

Besides the triplet ranking loss, another influential training criterion with negative samples is from the skip-gram model with negative sampling (Mikolov et al., 2013), where true context words c and randomly sampled negatives c_n enter the binary cross-entropy loss to encourage the similarity

of embeddings of target words x with similar contexts. The loss is characterized with the formula

$$\mathcal{L} = -\log \text{logistic}(c^T x) - \sum_n \log \text{logistic}(-c_n^T x)\,,$$

where $\text{logistic}(x) = \frac{1}{1+\exp(-x)}$. Such formulation of negative sampling with logistic activation is a simplification of the noise-contrastive estimation (Gutmann and Hyvärinen, 2010), and is a way to approximate and factorize the pointwise mutual information matrix (Levy et al., 2015), and can be seen as a softmax alternative to the triplet ranking loss (Dyer, 2014).

In this study, I use the triplet ranking loss with negative sampling to train the extraction module to generate close sentence representations for parallel sentences. I search over some small numbers of negative samples to draw randomly during training, but always include the hardest sample found in the mini-batch in light of Faghri et al. (2018)'s findings.

2.3 Adversarial networks for representation learning of texts

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 that representations "translated" to another language then "translated" back should be similar to the starting representations.

Conneau et al. (2018a)'s MUSE project trains a linear mapping from the word-embedding space of one language to that of another, using generative adversarial networks (GANs) (Goodfellow et al., 2014). The benefit of learning such alignment with GANs is that no parallel corpora are needed as the model learns to align the distributions of samples from different languages. This adversarial loss is later used for aligning translation sentences for an unsupervised machine translation model (Lample et al., 2018a).

Cycle consistency in complement to adversarial training (Zhu et al., 2017) 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).

Zhang et al. (2017) show the potential of unsupervised word translation with a model similar to cycle GAN (Zhu et al., 2017). In particular, they propose an unsupervised model where the generator transforms source embeddings to target space to make them indistinguishable by the discriminator and back-transforms them to the source space, where such back-translated embeddings are encouraged to be close to the original embedding by the reconstruction or cycle loss.

Xu et al. (2018) use the Sinkhorn distance, a distributional similarity measure, in place of the adversarial loss. They optimize linear maps in both directions (source to target and the reverse) for every language pair so that the source embeddings mapped to the target space match the distribution of the target embeddings. And when the mapped source embeddings from the target space are back-translated to the source space, they are trained to be maximally close to the original source embeddings, or with cycle consistency (Zhu et al., 2017).

Similar to the previous work, Mohiuddin and Joty (2020) also incorporate cycle consistency along with the adversarial loss to train the adversarial autoencoder. Contrary to the previous work, their model learns the mapping in a latent code space, instead of in the original embedding space. They also utilize and combine many refinement methods to improve the word embeddings learned from the adversarial autoencoder with cycle consistency. The system yields state-of-the-art performance on lexicon induction.

Romanov et al. (2019) explores domain disentanglement as well as alignment using WGANs (Arjovsky et al., 2017). Non-aligned sentences from corpora in two different domains are encoded, and a set of generator and discriminator is used to learn to produce embeddings which are partitioned into domain-general representations and domain-specific ones based on the distribution of data. The trained encoder is then evaluated for style-transfer. The benefit of using GANs is again that parallel sentences are not used, and therefore the representation learning is unsupervised.

This thesis work is the first to our knowledge to apply a strategy of adversarial training and cycle consistency to the task of bitext mining. And in this work, I adapt the architecture of Romanov et al. (2019) for the extraction module to learn cross-lingual representations based on monolingual corpora, as an alignment method of the unsupervised model.

2.4 Cross-lingual alignment in pretrained language models

This work adopts 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 learning cross-lingual transformation (Aldarmaki and Diab, 2019) or fine-tuning the pretrained model for higher alignment (Cao et al., 2020) with supervision from aligned texts.

In particular, Aldarmaki and Diab (2019) trains a linear mapping of sentence embeddings extracted from ELMo between languages using parallel sentences. They show that sentence representations obtained with this sentence-level mapping have better performance in the translation retrieval task than those with traditional word-level mapping. Their results suggest that the model can learn better cross-lingual sentence representation by being trained to align sentences rather than words.

And instead of training a linear mapping, Cao et al. (2020) fine-tune BERT by optimizing similarity of translation words in sentences, and show that the fine-tuning approach outperforms the linear mapping approach on zero-shot cross-lingual transfer of NLI benchmarks. This result suggests the possibilities that a more sophisticated aligner can be used to help generate sentence representations which are more similar to their translations in other languages.

Based on the findings from the previous research, I propose to train an extraction module which is more powerful than a linear mapping, and which learns to align sentence rather than word representations, in order to obtain a sentence encoder more suitable for cross-lingual transfers as well as translation retrievals. And this work follows this line of investigation into alternative methods to extract multilingual information from a language model pretrained with monolingual corpora.

2.5 Unsupervised parallel sentence mining

The evaluation task of this work is bitext mining without supervision from any bitexts or bitexts of the particular pair of languages, where bitext mining is the task of mining for the translation sentences in a corpus of target language given sentences in the source language. The experiments of unsupervised bitext mining have been explored previously.

Hangya et al. (2018) show that unsupervised bilingual word embeddings induced from adversarial autoencoder are effective on bitext mining, and Hangya and Fraser (2019) further improve the system with a cross-lingual word-alignment algorithm.

Kiros (2020) trains a lensing module over mBERT for the task of natural language inference (NLI), in the similar setup to Reimers and Gurevych (2019), and transfers the model to bitext mining. He also shows that the thresholds optimized for different langauge pairs are very close, which opens up the possibility of *threshold transfer* for bitext mining, the phenomenon of which this work analyzes and corroborates in Chapter 5.

Keung et al. (2020)'s unsupervised system based on mBERT (Devlin et al., 2019) bootstraps bitexts, and iteratively improves the retrieval capacity by fine-tuning the model using bootstrapped bitexts. Kvapilíková et al. (2020)'s system, on the other hand, synthesizes bitexts using an unsupervised machine translation system, and uses the synthesized translation pairs to fine-tune XLM (Conneau and Lample, 2019).

Results from the three aforementioned studies are included in Chapter 4 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 the model receives training signals from existing monolingual corpora or bitexts, instead of from NLI, bootstrapped, or synthesized data.

Chapter 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 with l layers, sequence length q, and embedding size d, the extraction module $f(\cdot)$ generates a sentence embedding y_i^γ with fixed size d based on the features x_i^γ, or

$$f(x_i^\gamma) = y_i^\gamma, \quad x_i^\gamma \in \mathbb{R}^{l \times q \times d} \text{ and } y_i^\gamma \in \mathbb{R}^d \,. \tag{3.1}$$

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 labeled with their language identities provide signals 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.

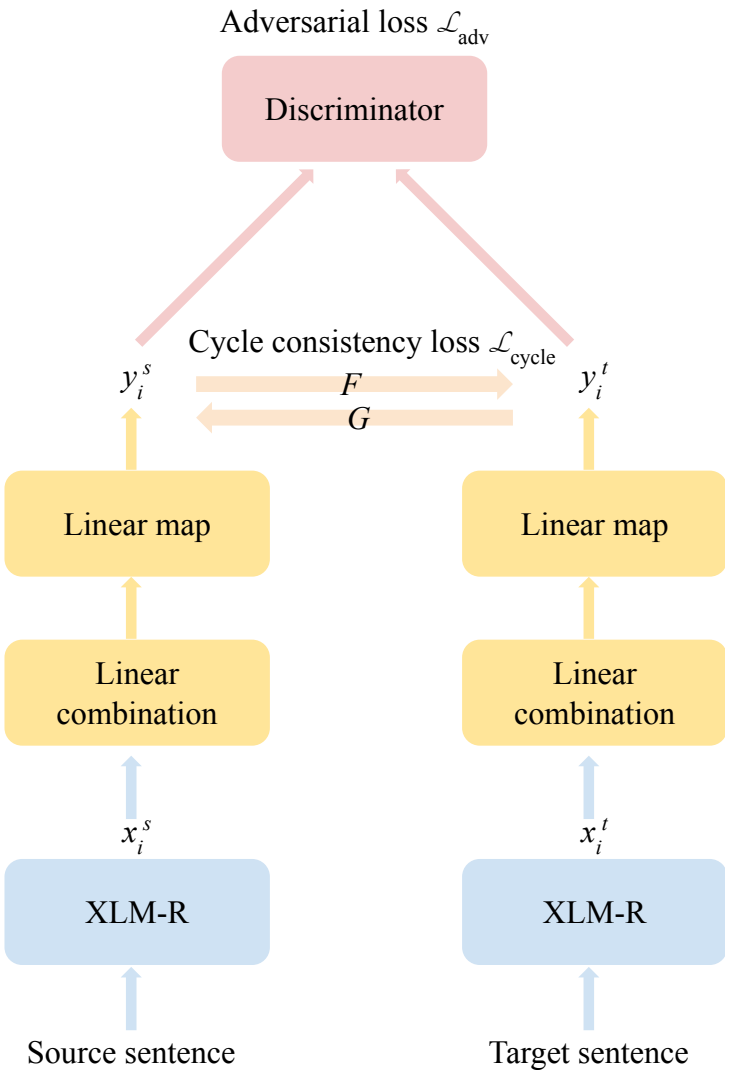

Figure 3.1: Schematic representation of the unsupervised model with the adversarial loss and the cycle consistency loss.

As is usual with GANs, there is a discriminator module $d(\cdot)$, which in this model consumes the representation y_i^γ and outputs 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

$$\mathcal{L}_{\text{disc}} = d(y_i^s) - d(y_j^t)\,. \tag{3.2}$$

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 so that the sentence encodings which are not translation pairs are dissimilar. 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:

$$\mathcal{L}_{\text{cycle}} = h(y_i^s, G(F(y_i^s))) + h(F(G(y_j^t)), y_j^t) \tag{3.3}$$

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

$$\begin{aligned} h(a,b) = \sum_n &\max(0, \alpha - \text{sim}(a,b) + \text{sim}(a_n,b)) \\ &+ \max(0, \alpha - \text{sim}(a,b) + \text{sim}(a,b_n))\,, \end{aligned} \tag{3.4}$$

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 are more effective than the sum of negatives in the ranking loss (Faghri et al., 2018), this 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 3.1.

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

$$\mathcal{L}_{\text{sup}} = h(y_i^s, y_i^t) \,, \tag{3.5}$$

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

3.4 Training

The model is trained by backpropagation with the Adam optimizer (Kingma and Ba, 2014), learning rate 0.001, and early stopping, with the parameters in XLM-R frozen. And the training program is built upon AllenNLP (Gardner et al., 2018), Transformers (Wolf et al., 2020), and PyTorch (Paszke et al., 2019).

For adversarial training, the discriminator is updated κ times for every step of backpropagation to the extraction module. Other hyperparameters include the dimension of the output representations d,

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 3.1: Hyperparameters and experimented values.

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 3.1. We empirically determine the hyperparameters among experimented values, and report their values in specific evaluation sections.

We experimented with two language pairs for training the model—Arabic-English (ar-en) and German-English (de-en)—to explore potential effects of the choice of the dual-pivots. The bilingual corpora used to train the model are accessed through the OPUS interface (Tiedemann, 2012), and the corpora are the same as the sentence-aligned corpora used to train the XLM (Conneau and Lample, 2019). The Arabic-English corpus is from the United Nations (UN) Parallel Corpus (Ziemski et al., 2016). The UN Parallel Corpus is comprised of manually translated UN documents in the six official languages of UN, and the fully sentence-aligned corpus comprises of 10 million parallel sentences. The German-English corpus, on the other hand, is from the European Union (EU) Bookshop Corpus (Skadiņš et al., 2014). The EU Bookshop Corpus is crawled from the EU Bookshop service, an online archive of publications from various EU institutions with manual translations in twenty-four official languages of EU. The German-English corpus consist of 9 million parallel sentences.

To train our model, raw texts of sentences are fed to the XLM-R (Conneau et al., 2020a), which segments the texts into subwords with its pretrained Byte-Pair Encoding (BPE) module (Sennrich et al., 2016). After being trained, the model is evaluated on two tasks of bitext mining, as described in the following chapter.

Chapter 4

EVALUATIONS

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

4.1 Baselines and comparisons

4.1.1 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.

Results from Kiros (2020), Keung et al. (2020), and Kvapilíková et al. (2020), as state-of-the-art models for unsupervised bitext mining based on pretrained language models, are also included for comparison (see Section 2.5 for a brief description of these models).

4.1.2 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 with 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 when the method of margin-based retrieval (Artetxe and Schwenk, 2019a) is not used.

4.2 BUCC

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

4.2.1 Margin-based retrieval

The retrieval is based on the margin-based similarity scores (Artetxe and Schwenk, 2019a),

$$\begin{aligned} \text{score}(y^s, y^t) &= \text{margin}(\text{sim}(y^s, y^t), \text{scale}(y^s, y^t)) \\ \text{scale}(y^s, y^t) &= \sum_{z \in \text{NN}_k(y^s)} \frac{\text{sim}(y^s, z)}{2k} + \sum_{z \in \text{NN}_k(y^t)} \frac{\text{sim}(y^t, z)}{2k}, \end{aligned} \tag{4.1}$$

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; Kiros, 2020; Keung et al., 2020). 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. The margin-based similarity scoring is a generalization of the cross-domain similarity local scaling (CSLS) proposed by Conneau et al. (2018a). Formula (4.1) with the difference margin function, or $\text{margin}(a, b) = a - b$, is the original formulation of CSLS.

Following Hu et al. (2020), the 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, so their results are not included in the table for this evaluation task.

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
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	89.8	92.5	88.7	88.9	88.9

Table 4.1: F1 scores on the BUCC training split. Simple average of scores from 4 tasks reported in the second column. Highest scores in their groups are bolded.

4.2.2 Results

F1 scores for the mining task with the BUCC dataset presented in Table 4.1 demonstrates 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 unsupervised model trained adversarially, outperforming the unsupervised baselines and reaching the state of the art, is 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 with the de-en texts performing better than that with the ar-en texts at mining for parallel sentences between English and German by 7 points. The results suggest that while alignment strategy is transferable, the unsupervised model can be further improved for multilingual alignment by being trained with multilingual texts with 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 difference 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.

4.3 Tatoeba

We also measure the parallel sentence matching accuracy over the Tatoeba dataset (Artetxe and Schwenk, 2019b). The Tatoeba dataset was constructed by Artetxe and Schwenk (2019b) from parallel sentences in the Tatoeba platform, a collaborative online database hosting example translation sentences for the purpose of foreign language learning for humans. This dataset consists of English-aligned sentence pairs for 112 languages. For 72 languages, there are 1,000 English-aligned sentences, and for the other 40 languages the size of bitext ranges from around 100 to around 900. The average size is of the bitexts if 794. A full list of the languages included in the dataset and the sizes of every parallel corpus can be found at Table A.1 in Appendix A. The task is to retrieve the translation pairs in the target language given those in the source language using absolute similarity scores without margin-scaling.

4.3.1 Results

Matching accuracy for the retrieval task with Tatoeba dataset are presented in Table 4.2. 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

Model	Average accuracy (%)		
	36	41	112
Unsupervised			
Model (ar-en unsup.)	73.4	70.8	55.3
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 (de-en bitexts sup.)	**80.4**	**78.9**	**62.4**
Fully Supervised			
LASER	85.4	79.1	66.9
LABSE	**95.0**	−−	**83.7**

Table 4.2: Average accuracy scores on the Tatoeba dataset in three average groups. Highest scores in their groups are bolded.

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. The choice of the pivot language pairs with this task does not show much difference.

Chapter 5

ANALYSIS

To understand the factors affecting the performance of the model, we consider ablated variants of the model. All models presented in this section are trained with German-English corpora with the same hyperparameters presented in the evaluation section above.

5.1 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). Here we study the phenomenon by examining the mining performance on the BUCC dataset of two threshold schemes: the optimized thresholds, where the thresholds are optimized for every language pair as in the evaluation section, 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 threshold schemes on the BUCC mining task are summarized in Table 5.1. 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 scores from the two schemes are almost identical with 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 with other language pairs.

5.2 Ablation

Here we ablate from the model the two trainable components, which are the weighted linear combination (Peters et al., 2018) and the linear map, as well as the two training losses, $\mathcal{L}_{\text{adv}}$ and

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 5.1: F1 scores on the BUCC training split with different threshold schemes. The models are trained with de-en texts. The system with optimized thresholds tunes the threshold for each of the four pairs, as in Table 4.1, whereas the dual-pivot threshold system adopts the threshold tuned from de-en pair for all four pairs.

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 5.2: Ablation results on the Tatoeba averaged across 36 languages in accuracy (%).

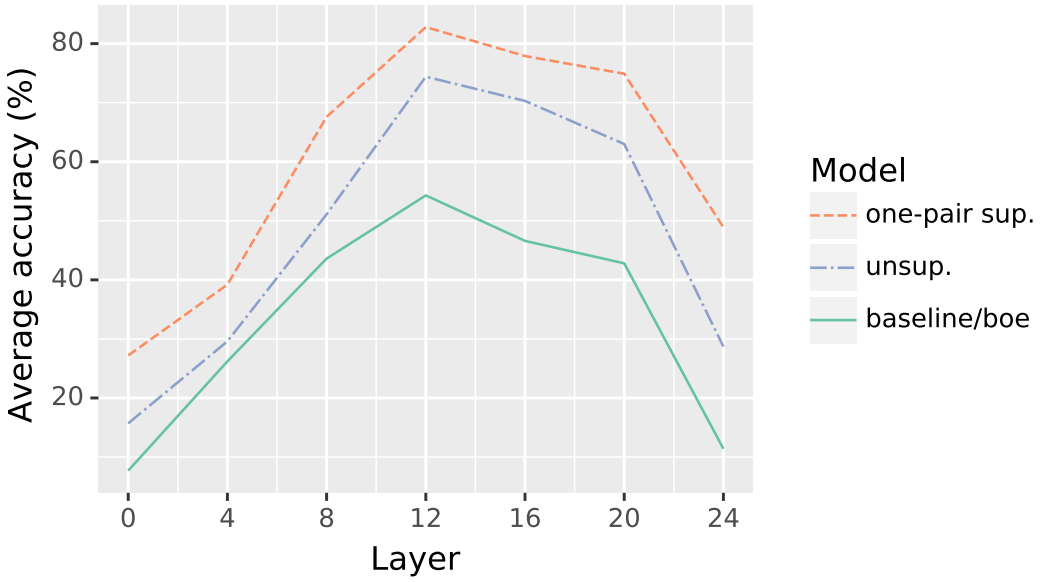

Figure 5.1: Average accuracies of 36 languages of the Tatoeba task with model trained with representations from XLM-R layers of orders which are multiples of 4.

$\mathcal{L}_{\text{cycle}}$, of the unsupervised model. When the weighted linear combination is ablated, the simple unweighted average of embeddings across layers is used instead.

The ablation results evaluated on the average accuracy over the 36 languages of the Tatoeba corpus and summarized in Table 5.2 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 the 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.

5.3 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). In order to understand this phenomenon in the present setting, we experiment with 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 5.1 show that the alignment information is most extractable deep in the middle layers, corroborating the findings from the previous work. 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.

5.4 Error analysis

When does the model make mistakes? Here I consider the Tatoeba retrieval task with the the language pair of Mandarin Chinese and English. While the full supervised model LASER (Artetxe and Schwenk, 2019b) reaches 95.9% accuracy with this pair, our unsupervised model gives 79.0% accuracy with it. Therefore, this pair represents a situation where the unsupervised model does well but still has sizable room of improvement. I inspect the roughly 200 errors made by the system, and categorize the errors based on factors which seem most salient and can perhaps cause the model to encode them with similar representations. The proportion of all error types are summarized in Table 5.3.

Half of the retrieval errors cannot be identified as belonging to a common category with any others, and are labeled as "unclassified" in the Table 5.3. The other half, however, demonstrates some interesting patterns. First, sharing the subjects, predicates, or adverbials accounts for about a third of the errors. This suggests that the model is trained to rely heavily on lexical items to produce cross-lingually aligned sentence embeddings. Secondly, sentences with matching sentiments or speech acts also seem to be encoded with similar embeddings in some error cases, which suggest

Error type	%	Example	True translation	Retrieved sentence
Modal match	2	熊會爬樹。	Bears can climb trees.	I can swim very fast.
Subject match	15	我在這裏出差。	I am here on business.	I am going to the store now.
Predicate similar	15	我睡不著。	I couldn't sleep.	You should sleep.
Adverbial match	2	我們今晚在外面吃飯吧。	Let's eat out tonight.	They set out last night.
Syntax similar	3	蛋黃是黃色的。	Yolks are yellow.	Pandas are beautiful animals.
Sentiment match	3	這可眞好。	This is quite good.	They love that.
Speech act match	6	請通知我。	Please keep me updated.	Let me think.
Reversing polarity	3	我不同意你這個觀點。	I can't go along with you on that point.	I agree with your opinion.
Missing information	1	我和一位朋友住在一起。	I am staying with a friend.	We live together.
Unclassified	49	法語課過得怎樣？	How was the French class?	Cheap meat doesn't make good soup.

Table 5.3: Proportions of the error categories and an example for each error category.

that the model may have picked up some pragmatic and or social information from the lexical clues. Finally, it is possible that the encoder does not encode much information about negation, and sentences with opposite truth conditions can be close in the joint embedding space, as is demonstrated by the retrieval errors of reversing polarity.

Chapter 6

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 model, and both unsupervised and one-pair supervised models shows that much 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 obtain sentence embeddings with higher multilingual alignment through the transferability of bilingual alignment demonstrated in this work, and such work will benefit tasks involving languages with low resources in bitexts.

BIBLIOGRAPHY

Aldarmaki, H. and Diab, M. (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.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of Machine Learning Research*, volume 70, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

Artetxe, M. and Schwenk, H. (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.

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

Cao, S., Kitaev, N., and Klein, D. (2020). Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (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.

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

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018a). Word translation without parallel data. *ArXiv:1710.04087 [cs.CL]*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018b). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (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.

Devlin, J., Chang, M.-W., Lee, K., and 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dyer, C. (2014). Notes on noise contrastive estimation and negative sampling. *arXiv:1410.8251 [cs.LG]*.

Faghri, F., Fleet, D. J., Kiros, J. R., and Fidler, S. (2018). VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT sentence embedding. *ArXiv:2007.01852 [cs.CL]*.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (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.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. JMLR Workshop and Conference Proceedings.

Hangya, V., Braune, F., Kalasouskaya, Y., and Fraser, A. (2018). Unsupervised parallel sentence extraction from comparable corpora. In *International Workshop on Spoken Language Translation*.

Hangya, V. and Fraser, A. (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.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine*

Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421, Virtual. PMLR.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Keung, P., Salazar, J., Lu, Y., and Smith, N. A. (2020). Unsupervised bitext mining and translation via self-trained contextual embeddings. *ArXiv:2010.07761 [cs.CL]*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv:1412.6980 [cs.LG]*.

Kiros, J. (2020). Contextual lensing of universal sentence representations. *ArXiv:2002.08866 [cs.CL]*.

Kiros, J. and Chan, W. (2018). InferLite: Simple universal sentence representations from natural language inference data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4868–4874, Brussels, Belgium. Association for Computational Linguistics.

Kvapilíková, I., Artetxe, M., Labaka, G., Agirre, E., and Bojar, O. (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.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018a). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (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.

Lazaridou, A., Dinu, G., and Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. Association for Computational Linguistics.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-*

nologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, M. E., Ruder, S., and Smith, N. A. (2019). To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Pires, T., Schlinger, E., and Garrette, D. (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.

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

Reimers, N. and Gurevych, I. (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.

Romanov, A., Rumshisky, A., Rogers, A., and Donahue, D. (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.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International*

Conference on Language Resources and Evaluation (LREC'14), pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tiedemann, J. (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).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xu, R., Yang, Y., Otani, N., and Wu, Y. (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.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (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.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Zweigenbaum, P., Sharoff, S., and Rapp, R. (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).

Appendix A

SUMMARY OF THE TATOEBA DATASET

ISO 639-3	Name	Size
afr	Afrikaans	1000
amh	Amharic	168
ang	Old English	134
ara	Arabic	1000
arq	Algerian Arabic	911
arz	Egyptian Arabic	477
ast	Asturian	127
awa	Awadhi	231
aze	Azerbaijani	1000
bel	Belarusian	1000
ben	Bengali	1000
ber	Berber	1000
bos	Bosnian	354
bre	Breton	1000
bul	Bulgarian	1000
cat	Catalan	1000
cbk	Chavacano	1000
ceb	Cebuano	600
ces	Czech	1000
cha	Chamorro	137

ISO 639-3	Name	Size
cmn	Mandarin Chinese	1000
cor	Cornish	1000
csb	Kashubian	253
cym	Welsh	575
dan	Danish	1000
deu	German	1000
dsb	Lower Sorbian	479
dtp	Kadazandusun	1000
ell	Modern Greek	1000
epo	Esperanto	1000
est	Estonian	1000
eus	Basque	1000
fao	Faroese	262
fin	Finnish	1000
fra	French	1000
fry	West Frisian	173
gla	Scottish Gaelic	829
gle	Irish	1000
glg	Galician	1000
gsw	Swiss German	117
heb	Modern Hebrew	1000
hin	Hindi	1000
hrv	Croatian	1000
hsb	Upper Sorbian	483
hun	Hungarian	1000
hye	Eastern Armenian	742

ISO 639-3	Name	Size
ido	Ido	1000
ile	Interlingue	1000
ina	Interlingua	1000
ind	Indonesian	1000
isl	Icelandic	1000
ita	Italian	1000
jav	Javanese	205
jpn	Japanese	1000
kab	Kabyle	1000
kat	Georgian	746
kaz	Kazakh	575
khm	Khmer	722
kor	Korean	1000
kur	Kurdish	410
kzj	Coastal Kadazan	1000
lat	Latin	1000
lfn	Lingua Franca Nova	1000
lit	Lithuanian	1000
lvs	Latvian	1000
mal	Malayalam	687
mar	Marathi	1000
max	North Moluccan	284
mhr	Meadow Mari	1000
mkd	Macedonian	1000
mon	Mongolian	440
nds	Low German	1000

ISO 639-3	Name	Size
nld	Dutch	1000
nno	Nynorsk	1000
nob	Bokmål	1000
nov	Novial	257
oci	Occitan	1000
orv	Old East Slavic	835
pam	Kapampangan	1000
pes	Western Persian	1000
pms	Piedmontese	525
pol	Polish	1000
por	Portuguese	1000
ron	Romanian	1000
rus	Russian	1000
slk	Slovak	1000
slv	Slovene	823
spa	Spanish	1000
sqi	Albanian	1000
srp	Serbian	1000
swe	Swedish	1000
swg	Swabian	112
swh	Swahili	390
tam	Tamil	307
tat	Tatar	1000
tel	Telugu	234
tgl	Tagalog	1000
tha	Thai	548

ISO 639-3	Name	Size
tuk	Turkmen	203
tur	Turkish	1000
tzl	Talossan	104
uig	Uyghur	1000
ukr	Ukrainian	1000
urd	Urdu	1000
uzb	Uzbek	428
vie	Vietnamese	1000
war	Waray	1000
wuu	Wu	1000
xho	Xhosa	142
yid	Yiddish	848
yue	Yue	1000
zsm	Malaysian	1000

Table A.1: Summary of the Tatoeba dataset. The table shows the names of the languages and the number of English-aligned sentences of every language in the Tatoeba dataset.