Enhancing Cross-lingual Sentence Embedding for Low-resource Languages with Word Alignment

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, Yoshimasa Tsuruoka
The University of Tokyo, Tokyo, Japan
{mzt, qiyuw, zhaokaiyan1006, zw2599, yoshimasa-tsuruoka}@g.ecc.u-tokyo.ac.jp

Abstract

The field of cross-lingual sentence embeddings has recently experienced significant advancements, but research concerning low-resource languages has lagged due to the scarcity of parallel corpora. This paper shows that cross-lingual word representation in low-resource languages is notably under-aligned with that in high-resource languages in current models. To address this, we introduce a novel framework that explicitly aligns words between English and eight low-resource languages, utilizing off-the-shelf word alignment models. This framework incorporates three primary training objectives: aligned word prediction and word translation ranking, along with the widely used translation ranking. We evaluate our approach through experiments on the bitext retrieval task, which demonstrate substantial improvements on sentence embeddings in low-resource languages. In addition, the competitive performance of the proposed model across a broader range of tasks in high-resource languages underscores its practicality.

1 Introduction

Cross-lingual sentence embedding encodes multilingual texts into a shared semantic embedding space in which the texts are understandable across different languages. Various applications including bitext retrieval (Artetxe and Schwenk, 2019a) and cross-lingual semantic textual similarity tasks (Cer et al., 2017; Chen et al., 2022) rely on cross-lingual sentence embedding.

Current approaches to obtaining cross-lingual sentence embeddings primarily utilize multilingual pre-trained language models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020) that employ masked language modeling and translation language modeling objectives to predict masked tokens within the context. Such models implicitly align the contextual representations of semantically similar units of sentences in different languages (Li et al., 2021), thereby enabling the models to understand texts in various languages.

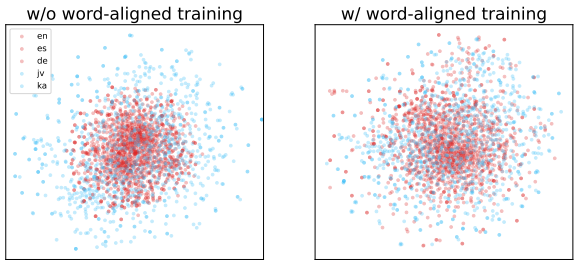

Figure 1: t-SNE visualization of sampled word embeddings from both high-resource and low-resource languages. The red points represent the word embeddings from high-resource languages, and the blue points correspond to those from low-resource languages. This comparison highlights the differences of word representation in the models w/ and w/o the explict word-aligned training. **Left**: words in low-resource languages are under-aligned with their translations in high-resource languages. **Right**: the phenomenon of under-alignment is mitigated through the proposed explicit word-aligned training. The details of word sampling, word embeddings and word-aligned training are described in Section 4.3.

While the field of cross-lingual sentence embedding has recently seen great advancements (Li et al., 2023, 2021; Zhang et al., 2023; Feng et al., 2022), research concerning low-resource languages has lagged due to the scarcity of parallel corpora.

In Figure 1, we observe that word embeddings from low-resource languages, which are derived from current cross-lingual models trained solely with a sentence-level alignment objective, are under-aligned with those from high-resource languages. To address this under-alignment, we introduce a new framework featuring two word-level alignment objectives: aligned word prediction and word translation ranking. These objectives are designed to align the word-level signals of parallel sentences. Additionally, a sentence-level alignment objective, known as translation ranking (Feng et al.,

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3225–3236

2022), is also used to ensure the basic sentence understanding. We name our proposed framework WACSE (**W**ord **A**ligned **C**ross-lingual **S**entence **E**mbedding). The right sub-figure in Figure 1 shows the distribution of word embeddings obtained from the model trained with the proposed aligned word prediction and word translation ranking. It demonstrates that the under-alignment phenomenon can be mitigated through the explicitly word-aligned objectives.

The experiment results demonstrate that the proposed word-aligned training objectives can enhance cross-lingual sentence embedding, particularly for low-resource languages, as evidenced on the Tatoeba dataset (Artetxe and Schwenk, 2019a). This finding matches our observations on word representations in Figure 1. Furthermore, our model retains competitive results across a broader range of tasks, including STS22 (Chen et al., 2022), BUCC (Zweigenbaum et al., 2017), and XNLI (Conneau et al., 2018), in which most languages are high-resource. This indicates the practicality and robustness of the proposed framework.

2 Related Work

2.1 Cross-lingual Sentence Embedding

Cross-lingual sentence embedding is the task of encoding sentences from various languages into a shared embedding space. Traditionally, large-scale parallel corpora have been utilized to learn cross-lingual sentence embeddings. LASER (Artetxe and Schwenk, 2019b) employs a BiLSTM encoder trained on parallel sentences from 93 languages, totaling 223 million parallel sentences, to learn joint multilingual sentence representations. LaBSE (Feng et al., 2022) learns cross-lingual sentence embeddings by integrating dual-encoder translation ranking, additive margin softmax, masked language modeling (MLM) and translation language modeling (TLM), utilizing training data consisting of 17 billion monolingual sentences and 6 billion translation pairs. Extending SimCSE (Gao et al., 2021) to multilingual settings, mSimCSE (Wang et al., 2022) demonstrates that contrastive learning applied to English data alone can yield universal cross-lingual sentence embeddings without the need for parallel data. Inspired by PCL (Wu et al., 2022), MPCL (Zhao et al., 2024) leverages multiple positives from different languages to improve cross-lingual sentence embedding.

Token-level auxiliary tasks. Recently, the importance of token-level auxiliary tasks has been recognized. VECO2.0 (Zhang et al., 2023) employs thesauruses for token-to-token alignment, achieving notable results on the XTREME benchmark (Hu et al., 2020). DAP (Li et al., 2023) is designed with two primary objectives. The first objective, translation ranking (TR), aims to bring parallel sentences closer together in the embedding space. The second objective, representation translation learning (RTL), employs one-sided contextualized token representations to reconstruct their translation counterparts, aiming to capture the relationships between tokens in parallel sentences. TR as a simple but effective objective, is also utilized in our framework to ensure the basic sentence understanding. Nevertheless, researchers recognize the significance of token-level or word-level alignment in cross-lingual scenarios, the acquisition of token-level or word-level supervisory signals remains a challenging topic of ongoing discussion. Li et al. (2021) employ fast_align (Dyer et al., 2013) to obtain word-level supervisory signals. XLM-Align (Chi et al., 2021b) leverages self-labeled word alignment signals for model training. VECO2.0 (Zhang et al., 2023) utilizes thesauruses to acquire token-level supervisory signals.

2.2 Word Alignment

Word alignment is a task aimed at aligning the corresponding words in parallel sentences (Brown et al., 1993; Och and Ney, 2003; Dyer et al., 2013; Dou and Neubig, 2021; Wu et al., 2023), serving as a useful component for applications such as machine translation (Li et al., 2019, 2022). SimAlign (Jalili Sabet et al., 2020) utilizes multilingual word embeddings for word alignment without relying on parallel data or dictionaries. Nagata et al. (2020) redefine the word alignment task as a cross-lingual span prediction problem and fine-tune mBERT with manually annotated word alignment data. WSPAlign (Wu et al., 2023) reduces the dependence on manually annotated data by creating a large-scale, weakly-supervised dataset for word alignment. By pre-training word aligners with weakly-supervised signals via span prediction, it achieves state-of-the-art performance across five word alignment datasets. In this work, we employ WSPAlign to obtain the word-level supervisory signals for training models.

3 Method

To enhance cross-lingual sentence embeddings of low-resource languages through explicit alignment of words, WACSE incorporates three tasks, translation ranking (TR), aligned word prediction (AWP) and word translation ranking (WTR) tasks. These tasks collectively aim to learn the cross-lingual sentence representations of parallel sentences. The framework is depicted in Figure 2.

Formally, we start with a parallel dataset $(\mathbb{X}, \mathbb{Y})$ in two languages and the i-th parallel sentence pair is denoted as (X_i, Y_i). X_i and Y_i can be represented as a sequence of words: $X_i = x_1, x_2, \ldots, x_{|X_i|}$ and $Y_i = y_1, y_2, \ldots, y_{|Y_i|}$, respectively where $|\cdot|$ denotes the length of the given sentence. After inputting a sentence into the model, we obtain the hidden representations from the last layer as follows:

$$h_{cls}^{X_i}, h_1^{X_i}, h_2^{X_i}, \ldots, h_{|X_i|}^{X_i} = f(X_i), \quad (1)$$

where f represents the encoder, and $h_i^{X_i}$ denotes the corresponding hidden representation of x_i in sentence X_i.

Note that $h_i^{X_i}$ could be a sequence of embeddings because a word could be tokenized into multiple tokens. This could affect some minor implementation in the practice. Refer to Section 4.3 for the detailed implementation regarding this issue. Particularly, $h_{cls}^{X_i}$ is the hidden state of the `cls` token for representing the whole sentence.

Acquisition of Word Alignment Supervision. Word alignment models enable us to identify semantically equivalent word-level units within parallel sentences. We utilize WSPAlign[1] to obtain the word-level supervisory signals which will be used in the calculation of AWP and WTR losses.

For the i-th parallel sentence (X_i, Y_i), a word alignment model can generate bidirectional word pair dictionary $\text{WA}^{X_i \to Y_i}$ and $\text{WA}^{Y_i \to X_i}$ as follows:

$$\text{WA}^{X_i \to Y_i}, \text{WA}^{Y_i \to X_i} = \text{WordAlign}(X_i, Y_i). \quad (2)$$

Using $\text{WA}^{X_i \to Y_i}$, we can look up an aligned word $y_k \in Y_i$ for a specific $x_j \in X_i$, if it exists, and vice versa. The bidirectional dictionaries record all obtainable word pairs, demonstrated by the following equation:

$$y_k = \text{WA}^{X_i \to Y_i}(x_j), 1 \le j \le |X_i|. \quad (3)$$

Here, each word pair (x_j, y_k) represents a semantically equivalent word pair from the two sentences. In practice, we exclude word pairs with alignment scores below a specified threshold. The threshold value[2] for WSPAlign which we use is set to 0.9.

3.1 Aligned Word Prediction (AWP) Task

After obtaining word alignment supervisory signals, we introduce AWP objective to align semantically equivalent words across different languages.

For a word pair (x_j, y_k) derived from (X_i, Y_i), as introduced in Equations 2 and 3, the model is tasked with predicting y_k while x_j is masked.

We define the aligned word prediction loss for X_i as follows:

$$l^{AWP}(X_i) = \sum_{x_j \in \text{WA}^{X_i \to Y_i}} \text{MLM}(X_i, x_j; y_k), \quad y_k = \text{WA}^{X_i \to Y_i}(x_j), \quad (4)$$

where masking language modeling (MLM) means that the model predicts y_k while masking x_j. The total loss of a batch $\mathcal{L}^{AWP}$ is given by:

$$\mathcal{L}^{AWP} = \frac{1}{2N} \sum_{(X_i, Y_i) \in (\mathbb{X}, \mathbb{Y})} (l^{AWP}(X_i) + l^{AWP}(Y_i)), \quad (5)$$

where N is the batch size. This calculation incorporates both $X_i \to Y_i$ and $Y_i \to X_i$ directions.

3.2 Word Translation Ranking (WTR) Task

Besides the AWP task, previous studies have shown that token-level contrastive learning is also effective in cross-lingual pre-training (Li et al., 2021; Zhang et al., 2023). Inspired by this, we introduce WTR task in this section. WTR differs from the approach taken by VECO2.0 (Zhang et al., 2023), which utilizes thesauruses for token-to-token contrastive learning. The thesaurus-based method overlooks the contextual information of parallel sentences. In contrast, our approach leverages word

[1] https://github.com/qiyuw/WSPAlign.InferEval

[2] https://github.com/qiyuw/WSPAlign.InferEval/blob/49ac6fb87fab17079153bcce84c3ac52d4ce6752/inference.py#L74C5-L74C24

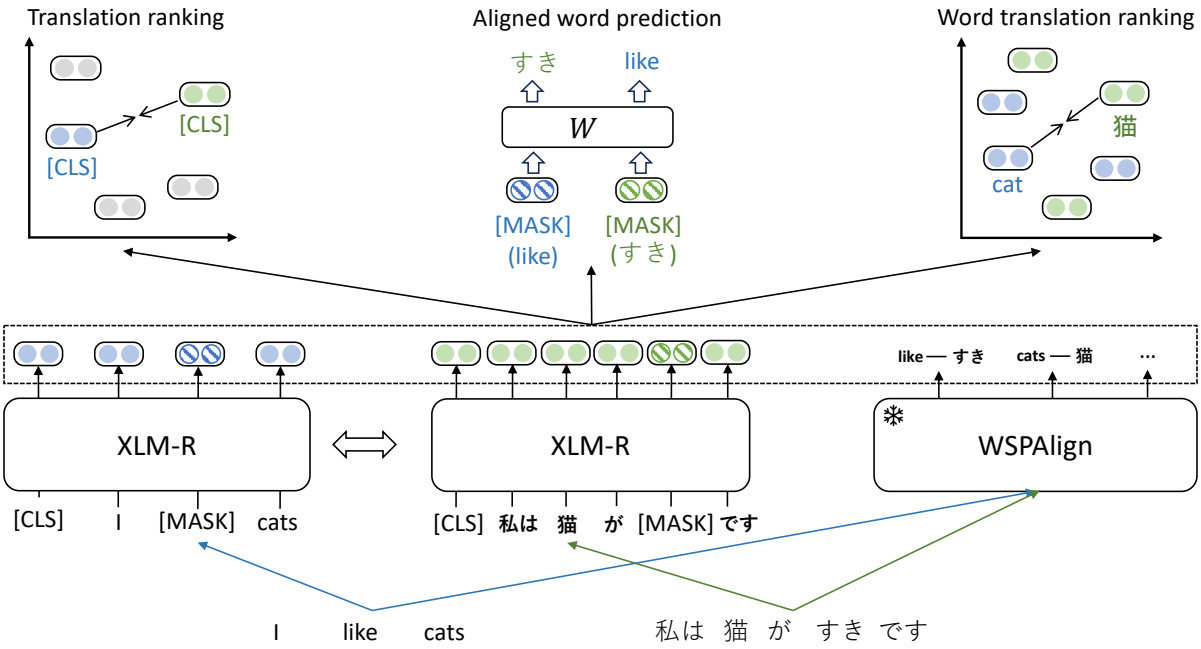

Figure 2: Illustration of WACSE framework. A parallel sentence pair is fed into the multilingual model along with a frozen word alignment model to obtain sentence representations, contextual token representations, and word alignment respectively. Then three objectives are calculated: (1) translation ranking: aligning sentence-level semantics; (2) aligned word prediction: utilizing the contextual representations of masked words to predict their aligned counterparts in another language; and (3) word translation ranking: aligning word-level semantics.

pair supervision from word alignment which considers the contextual information of words in parallel sentences to align semantically equivalent tokens within parallel sentences.

For a given sentence pair (X_i, Y_i) and a specific word pair (x_j, y_k) obtained from it, the word-level WTR loss $l^{WTR}(x_j)$ for x_j can be calculated as follows:

$$-\log \frac{e^{\phi^m(h_j^{X_i}, h_k^{Y_i}))}}{e^{\phi^m(h_j^{X_i}, h_k^{Y_i})} + \sum_{n=1 \wedge n \neq k}^{|Y_i|} e^{\phi^m(h_j^{X_i}, h_n^{Y_i})}}, \tag{6}$$

where ϕ^m particularly denotes a pair-wise cosine similarity function as the length of $h_j^{X_i}$ may not be equal to that of $h_k^{Y_i}$. Given that the word alignment model produces multiple word pairs, the loss for the whole sentence X_i is calculated as:

$$l^{WTR}(X_i) = \sum_{x_j \in \mathrm{WA}^{X_i \to Y_i}} l^{WTR}(x_j). \tag{7}$$

Considering bidirectional prediction across the entire batch, the loss $\mathcal{L}^{WTR}$ is presented as follows:

$$\frac{1}{2N} \sum_{(X_i, Y_i) \in (\mathbb{X}, \mathbb{Y})} (l^{WTR}(X_i) + l^{WTR}(Y_i)). \tag{8}$$

3.3 Translation Ranking (TR) Task

The dual-encoder architecture, combined with the TR task, has been shown to be effective in learning cross-lingual sentence embeddings at the sentence level, as evidenced by various studies (Guo et al., 2018; Yang et al., 2019; Feng et al., 2022). The TR task aligns the sentence representations of different languages at the sentence level to ensure the basic sentence understanding.

Following Feng et al. (2022) and Li et al. (2023), we denote the loss of the TR task for a parallel sentence (X_i, Y_i) as follows:

$$l_i^{TR} = -\log \frac{e^{\phi(h_{cls}^{X_i}, h_{cls}^{Y_i}))}}{e^{\phi(h_{cls}^{X_i}, h_{cls}^{Y_i})} + \sum_{j=1 \wedge j \neq i}^{N} e^{\phi(h_{cls}^{X_i}, h_{cls}^{Y_j})}}. \tag{9}$$

For the entire batch, the total loss of TR is:

$$\mathcal{L}^{TR} = \frac{1}{N} \sum_{i=1}^{N} l_i^{TR}, \tag{10}$$

where N represents the batch size, and ϕ denotes the cosine similarity function. The `cls` representations, $h_{cls}^{X_i}$ and $h_{cls}^{Y_i}$, are used to calculate the similarity between X_i and Y_i.

The final loss is calculated as the weighted sum

of three losses:

$$\mathcal{L} = \alpha\mathcal{L}^{TR} + \beta\mathcal{L}^{AWP} + \gamma\mathcal{L}^{WTR}, \quad (11)$$

where α is the weight for the TR loss, β for the AWP loss, and γ for the WTR loss.

4 Experimental Setup

4.1 Training Data

We utilize the same parallel corpora for training as DAP (Li et al., 2023), which is English-centric and comprises 36 language pairs. We use ISO 639 language codes[3] (two-letter codes) to denote languages. Using the same dataset as DAP, we employ WSPAlign to identify word-level semantically equivalent units. The statistics of the parallel corpora we use are presented in Table 1.

Language Pair	# Parallel Sentences	
	train	dev
en-kk	18190	2021
en-te	78105	8678
en-ka	146905	10K
en-jv	317252	10K
en-other	1M	10K

Table 1: Number of parallel sentences per language in the training and development corpora.

Lang. Code	# Articles	Lang. Code	# Articles
tl	45750	jv	72851
sw	78915	ml	84939
te	88914	mr	94005
af	113208	bn	144218
hi	159888	th	160499
ta	160712	ka	169878
ur	198346	el	228223
kk	235611	et	241085
bg	294740	he	345544
eu	424058	hu	533933
tr	540433	fi	563464
ko	652657	id	673857
fa	983682	pt	1114362
ar	1223016	vi	1289408
zh	1390659	ja	1395361
it	1838179	es	1911915
ru	1950729	nl	2141291
fr	2573743	de	2859124

Table 2: Number of Wikipedia articles available per language in the 36 languages of the training parallel corpora (accessed time: 2023-12-05 17:21:49).

[3] https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes

4.2 Low-resource Languages

We focus on cross-lingual sentence embeddings in low-resource languages. In the experiment, we determine low-resource languages based on two criteria: (1) the number of Wikipedia articles available per language[4] and (2) the size of the training data available for each language. Among the 36 languages in our dataset, six languages (tl, jv, sw, ml, te, mr) are identified as low-resource based on the smallest number of Wikipedia articles, according to criterion 1. For criterion 2, we select four languages (kk, te, ka, jv) with the fewest parallel sentences in the training set. Detailed information on the number of Wikipedia articles per language is available in Table 2. Considering the intersection of two criteria, we classify eight languages as low-resource in this study. Furthermore, we assess our proposed approach using various combinations of these eight languages, including settings with four languages (kk, te, ka, jv), five languages (tl, kk, te, ka, jv) and all eight languages.

4.3 Implementation Details

As we mentioned above, $h_j^{X_j}$ and $h_k^{Y_k}$ could consist of multiple hidden states and the number of them could be different. When calculating MLM loss in Equation 4, we roughly clip the longer sequence of tokens to ensure the number of tokens are equivalent.

As for the details of Figure 1, the words are sampled from the training dataset based on their frequencies, totaling 500 words. These word embeddings are extracted from the embedding layers of XLM-R models. The left sub-figure illustrates the result from the model trained solely with the TR objective, while the right sub-figure displays the result of our model, which is trained using the proposed method.

Model Size. For the Transformer encoder model (Vaswani et al., 2017) that we use, we adopt the configuration of XLM-R model (Conneau et al., 2020). We initialize the encoder model using the `xlm-roberta-base` checkpoint[5].

Hyperparameters. The maximum sequence length is set to 32. We train our model using the AdamW optimizer, with a learning rate of 5e-5. The training steps are 10K or 100K depending on

[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[5] https://huggingface.co/xlm-roberta-base

Model	4 langs	5 langs	8 langs	36 langs
LaBSE	92.5	93.5	93.8	95.4
InfoXLM	35.4	32.8	39.3	57.0
DAP	73.9	73.4	79.6	92.0
WACSE **(ours)**	**75.9**(+2.0)	**76.0**(+2.6)	**81.2**(+1.6)	**92.1**(+0.1)

Table 3: Average accuracy on the Tatoeba dataset across both directions for selected languages. The chosen low-resource languages are (kk, te, ka, jv) for "4 langs", (tl, jv, ka, kk, te) for "5 langs", and (te, ka, kk, jv, ml, sw, tl, mr) for "8 langs". Results of DAP are from Li et al. (2023).

different evaluation tasks. Gradient accumulation is employed across two A100 GPUs, resulting in a total batch size of 1024. The reported results are the average of two random seeds (42 and 0). The values of α, β and γ are set to 0.8, 0.1 and 0.1 empirically. For all models, the pooling method is configured as `cls_before_pooler`.

In line with DAP, we evaluate the model every 2,000 steps using development set shown in Table 1. Similarity search, which is a widely-used metric in cross-lingual retrieval tasks (Artetxe and Schwenk, 2019a), is utilized for choosing the optimal checkpoint.

4.4 Baselines

We compare our proposed method with XLM-R and its TR fine-tuned variant. Other competitive models such as InfoXLM (Chi et al., 2021a), LaBSE (Feng et al., 2022), and mSimCSE (Wang et al., 2022) are also included in the comparison. Note that some of these models leverage significantly larger datasets than ours. For instance, LaBSE utilizes 17 billion monolingual sentences and 6 billion translation pairs, while ours is only in the scale of 36 million.

Our main baseline is DAP (Li et al., 2023), which is a recent cross-lingual sentence embedding model leveraging token-level information. Hence, We adopt the identical settings including training data, model size, and other hyperparameters[6].

5 Evaluation Tasks and Results

5.1 Bitext Retrieval

Bitext retrieval is the task of retrieving the most relevant sentence from a target language corpus given a query sentence in the source language (Li et al., 2023). The Tatoeba dataset (Artetxe and Schwenk, 2019a) is a benchmark for evaluating bitext retrieval spanning a broad array of languages. We train our model for 100K steps and evaluate it on Tatoeba in this task. The released checkpoints of LaBSE[7] and InfoXLM[8] are used for comparison.

Results. We report the results of the Tatoeba dataset across four settings, as detailed in Table 3. The low-resource language settings, including the four-language, five-language, and eight-language settings, are described in Section 4.2. The thirty-six-language setting encompasses all 36 languages in the training dataset. From Table 3, we can observe that our model improves the cross-lingual sentence embedding in all low-resource language settings. But when expanding to all 36 languages, the improvement becomes marginal.

A possible explanation for this is that current cross-lingual sentence embedding models may struggle with learning the word-level alignment in low-resource languages due to the limited training data available. Through the explicit word-level alignment objectives, our method facilitates the alignment of the semantically equivalent tokens between high-resource languages and low-resource languages, aiding the model in acquiring basic word-level semantic information for low-resource languages. Therefore, The proposed method can improve cross-lingual sentence embeddings of low-resource languages. In contrast, high-resource languages already achieve effective word-level alignment during the pre-training phase with implicit word-level signals in the rich parallel corpus. Hence, continuing to explicitly align word-level semantic units between two high-resource languages could detract from the language-dependent and sentence-level features of the cross-lingual sentence embeddings.

[6] https://github.com/ChillingDream/DAP

[7] https://huggingface.co/sentence-transformers/LaBSE

[8] https://huggingface.co/microsoft/infoxlm-base

5.2 Cross-lingual Semantic Textual Similarity

Semantic Textual Similarity (STS) assesses the degree of similarity between two sentences. The cross-lingual STS task expands this to multilingual scenarios. For this task, we utilize STS22 (Chen et al., 2022) dataset and evaluate the performance using the MTEB benchmark (version 1.1.1) (Muennighoff et al., 2023). According to MTEB, the Spearman correlation, based on similarity, is the chosen metric for evaluation (Reimers et al., 2016). For both DAP and our method, we train the model for 10K steps and test on STS22. Zhao et al. (2024) point out that the result for $fr \leftrightarrow pl$ (French-Polish) language pair in STS22 seems unstable. Consequently, we report two versions of the STS22 task results, one including all language pairs in the STS22 dataset of MTEB benchmark and the other excluding the $fr \leftrightarrow pl$ pair.

Results. As shown in Table 4, our method significantly outperforms DAP on the STS22 dataset. This improvement illustrates that leveraging word-level semantically equivalent units, obtained through word alignment, can enhance the performance of cross-lingual sentence embedding models on cross-lingual STS tasks. This enhancement occurs by bringing semantically equivalent units closer across languages, even though the languages in STS22 are not considered low-resource. It is noteworthy that LaBSE performs slightly better than WACSE on STS22. Though LaBSE utilizes a much larger training dataset, WACSE still achieves competitive results. Detailed scores of STS22 are provided in Table 5.

Model	STS22	
	Avg.	Avg.(-fr-pl)
LaBSE	59.2	59.1
InfoXLM	49.6	47.5
XLM-R+DAP	51.7	52.1
XLM-R+ WACSE	**58.7(+7.0)**	**58.5(+6.4)**

Table 4: Spearman correlation scores of STS22. The results of LaBSE and InfoXLM are obtained using the MTEB benchmark.

5.3 Bitext Mining

Bitext mining involves extracting parallel sentences from two monolingual corpora with the assumption that some of these sentences are translation pairs. Following the settings of DAP and mSimCSE, we assess our model using the BUCC dataset (Zweigenbaum et al., 2017) which includes four language pairs: $fr \leftrightarrow en$, $de \leftrightarrow en$, $ru \leftrightarrow en$ and $zh \leftrightarrow en$. We train our model for 10K steps and use the evaluation code from mSimCSE[9].

	ar	de	de-en	de-fr	de-pl	en
XLM-R + DAP	49.2	38.0	43.3	49.8	43.6	55.2
XLM-R + Ours	55.2	41.6	47.9	52.2	49.9	60.2
	es	es-en	es-it	fr	fr-pl	it
XLM-R + DAP	59.0	62.1	55.1	67.1	45.0	66.2
XLM-R + Ours	60.0	70.0	65.5	73.4	62.0	71.2
	pl	pl-en	ru	tr	zh	zh-en
XLM-R + DAP	30.0	55.1	49.8	50.0	58.3	54.6
XLM-R + Ours	33.8	69.1	55.0	57.7	63.2	68.9

Table 5: Detailed results of the STS22 dataset.

Results. Table 6 shows the results of different models which we compare. The results of LASER (Artetxe and Schwenk, 2019a), mSimCSE, XLM-R, and LaBSE are from the mSimCSE paper (Wang et al., 2022). The notation "mSimCSE$_{sw,fr}$+NLI" refers to the variant of mSimCSE trained with a combination of English Natural Language Inference (NLI) data and translation pairs in English-Swahili and English-French (Wang et al., 2022). Our proposed method outperforms the "mSimCSE$_{sw,fr}$+NLI" model, even it is a large size model. From Table 6, we can see that our approach achieves competitive results, positioning it between the performance of "mBERT + DAP" and "XLM-R + DAP" at the base model size.

5.4 Cross-lingual Natural Language Inference

The Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018)) is a task that requires the model to classify sentence pairs across 15 languages into categories of entailment, neutrality, and contradiction. Following the settings of Chi et al. (2021a) and Li et al. (2023), we apply a cross-lingual transfer approach where the model is fine-tuned on English training data and then evaluated on test datasets in other languages. We use the same hyperparameter setting as DAP, with a batch size of 256 and a maximum sequence length of 128 tokens. The number of epochs is set to 2. We do not employ weight decay and experiment with learning rates of {1e-5, 3e-5, 5e-5, 7e-5}. The optimal learning rate is 7e-5 for our model.

Results. Table 7 shows the accuracy results. The XNLI task does not inherently depend on cross-lingual sentence embedding, thus not directly bene-

[9] https://github.com/yaushian/mSimCSE

Model	fr-en			de-en			ru-en			zh-en			avg.
	P	R	F	P	R	F	P	R	F	P	R	F	F
LASER	–	–	–	–	–	–	–	–	–	–	–	–	92.9
XLM-R large													
mSimCSE$_{sw,fr}$ + NLI	–	–	–	–	–	–	–	–	–	–	–	–	93.6
XLM-R base													
XLM-R	–	–	–	–	–	–	–	–	–	–	–	–	66.0
LaBSE	–	–	–	–	–	–	–	–	–	–	–	–	93.5
mBERT + DAP	94.1	92.9	93.5	97.5	93.8	95.6	96.7	90.8	93.7	94.5	93.2	93.8	94.1
XLM-R + DAP	94.1	93.2	93.7	97.5	95.6	96.5	97.8	94.2	96.0	96.4	93.6	95.0	**95.3**
XLM-R + WACSE **(ours)**	93.8	93.4	93.6	97.9	94.9	96.4	97.0	94.0	95.4	94.1	95.3	94.7	95.0(-0.3)

Table 6: Performance on the BUCC dataset. "mBERT + DAP" and "XLM-R + DAP" (Li et al., 2023) are our re-implemented results with the same 10K training steps as "XLM-R + WACSE". The results of LaBSE are from mSimCSE paper (Wang et al., 2022).

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg.
InfoXLM	86.4	80.3	80.9	79.3	77.8	79.3	77.6	75.6	74.2	77.1	74.6	77.0	72.2	67.5	67.3	76.5
LaBSE	85.4	80.2	80.5	78.8	78.6	80.1	77.5	75.1	75.0	76.5	69.0	75.8	71.9	71.5	68.1	76.3
XLM-R	83.8	77.6	78.2	75.4	75.0	77.0	74.8	72.7	72.0	74.5	72.1	72.9	69.6	64.2	66.0	73.7
+ TR	83.5	76.4	76.8	75.7	74.2	76.2	74.6	71.8	71.1	74.2	69.1	72.9	68.8	66.8	65.2	73.1
+ TR + TLM	84.6	77.4	76.9	74.9	68.1	69.8	69.4	68.1	61.7	68.9	62.6	66.9	61.4	61.7	57.5	68.7
+ DAP	82.9	77.0	77.7	75.7	75.2	76.0	74.7	73.1	72.5	74.2	71.9	73.0	69.8	70.5	66.0	74.0
+ WACSE **(ours)**	83.8	77.7	78.2	76.5	75.4	77.2	75.0	73.1	72.1	74.9	72.3	73.5	69.9	69.0	65.0	**74.2(+0.2)**

Table 7: Accuracy on the XNLI dataset. Results of DAP (Li et al., 2023), InfoXLM (Chi et al., 2021a) and LaBSE (Feng et al., 2022) are taken from DAP paper (Li et al., 2023).

fiting from the training in a straightforward manner, but our model demonstrates a slight improvement over the DAP model. Unlike DAP, which utilizes the Representation Translation Learning (RTL) objective to understand token-level relationships between parallel sentences, our model employs a novel framework with two word-level alignment objectives to align semantically equivalent token representations across languages. This suggests that our framework's approach may offer marginal advantages over the RTL loss used by DAP in capturing the nuanced semantics necessary for cross-lingual natural language inference. Note that this task does not directly pertain to cross-lingual sentence embedding. As a result, this observation also illustrate the practicality of our framework.

6 Analysis

In this section, we carry out experiments to gain a deeper understanding of the proposed framework, specifically investigating the role of language identification information and the three losses in WACSE. Our primary focus is on the Tatoeba dataset as it is the only one that encompasses the low-resource language setting. All models discussed in this section are trained with a fixed seed (42) and the training step is 10K.

6.1 Does the Language Identification Information Matter?

We conduct experiments to determine whether incorporating language-specific information can enhance cross-lingual sentence embeddings of low-resource languages. Specifically, we add a new embedding layer that encodes language IDs as the language embedding. We assign different ID number for different languages. It is then added to the token embeddings of our models. This approach is designed to assess the significance of language identification information for our method. We initialize the language embedding layer randomly at the start of training. The final embedding fed into the models is the sum of the token embedding, the positional embedding and the language embedding.

According to the results presented in Table 8, for low-resource languages, incorporating language identification information proves to be beneficial. However, for the cross-lingual sentence embeddings of all 36 languages, it appears more advan-

tageous not to include the language identification information.

Model	5 langs	36 langs
WACSE (w/o lang embed)	77.4	91.1
WACSE (w/ lang embed)	78.2	90.8

Table 8: Average accuracy across two directions on the Tatoeba dataset for five low-resource languages and all 36 languages. "w/ lang embed" denotes models trained with the language embedding layer, while "w/o lang embed" refers to models without this layer.

6.2 Do Word-level Objectives Matter?

We also train models exclusively on the TR task to highlight the effectiveness of the AWP and WTR objectives. Following Section 6.1, we present results for both the low-resource language setting and the 36-language setting. As indicated in Table 9, the AWP and WTR objectives prove to be effective in both scenarios. Note that their performance in the low-resource language setting surpasses that in the 36-language setting.

Model	5 langs	36 langs
TR	75.8	90.7
WACSE	77.4	91.1

Table 9: Average accuracy for the both directions on the Tatoeba dataset across five low-resource languages and all 36 languages. "TR" represents the model trained solely with the translation ranking objective, while WACSE refers to the model trained with the TR, AWP and WTR objectives.

6.3 Can AWP and WTR be Used Solely?

We present the results of models trained with the TR and AWP objectives, the TR and WTR objectives and a combination of the three objectives (TR, AWP and WTR). To accurately investigate the effect of AWP and WTR, we conduct grid search to find the optimal hyperparameters for the model trained with the combined three objectives. Specifically, the loss weights are 0.8, 0.02 and 0.18, for TR, AWP and WTR in the WACSE in Table 10, respectively.

As illustrated in Table 10, both AWP and WTR contributes to enhancing the cross-lingual sentence embeddings for low-resource languages in comparison to the model utilizing only the TR objective. Moreover, WTR exhibits a marginally superior capability for learning cross-lingual sentence embeddings than AWP. The advantage may stem from WTR's strategy of aligning word-level equivalent units within the context of parallel sentences, whereas AWP focuses on predicting masked tokens using the context of monolingual sentences. The optimal result is the combination of three objectives, showing the effectiveness of our WACSE framework.

Model	8 langs
TR	79.8
TR + AWP	80.8
TR + WTR	81.1
WACSE	81.2

Table 10: Average accuracy across both directions on the Tatoeba benchmark dataset for the eight low-resource language setting. "TR" indicates the model trained exclusively with the translation ranking (TR) objective. "TR + AWP" refers to the model trained with both the TR and AWP objectives. "TR + WTR" represents the model trained with the TR objective and the WTR objective. WACSE denotes the model trained with a combination of the TR, AWP and WTR objectives.

7 Conclusion

In this paper, we observe an intriguing phenomenon: the distributions of word embeddings of low-resource languages are under-aligned with those of high-resource languages in current multilingual pre-trained language models. Based on this observation, we propose a framework designed to align word-level semantically equivalent units in parallel sentences between high-resource languages and low-resource languages, thereby enhancing the cross-lingual sentence embeddings for low-resource languages. Furthermore, we demonstrate that aligning word-level semantically units between two high-resource languages with our proposed method may detrimentally affect the language-specific features learned during the pre-training phase. Our experimental results show the effectiveness of our method in improving cross-lingual sentence embeddings for low-resource languages. Additionally, WACSE preserves the performance of the model on other tasks that involve high-resource languages.

Limitations

Our approach does not consider phrase-level alignment between high-resource languages and low-resource languages, an aspect that merits further investigation. The effectiveness of our proposed method is significantly influenced by the quality of the word alignment model, i.e., WSPAlign. The released WSPAlign was not trained for low-resource languages particularly. Thus, developing a word alignment model with strong cross-lingual transferability is an important future direction.

Ethics Statement

All datasets and checkpoints used in this paper are copyright free for research purpose. Previous studies are properly cited and discussed. This research aims to improve cross-lingual sentence embedding models for low-resource languages. We do not introduce additional bias to particular communities. We utilized LLM only for proofreading but not generating any specific contents in this paper.

References

Mikel Artetxe and Holger Schwenk. 2019a. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. 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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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 *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. 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 generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Lei Li, Kai Fan, Hongjia Li, and Chun Yuan. 2022. Structural supervision for word alignment and machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4084–4094, Dublin, Ireland. Association for Computational Linguistics.

Shicheng Li, Pengcheng Yang, Fuli Luo, and Jun Xie. 2021. Multi-granularity contrasting for cross-lingual pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1708–1717, Online. Association for Computational Linguistics.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Dual-alignment pre-training for cross-lingual sentence embedding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3466–3478, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11084–11099, Toronto, Canada. Association for Computational Linguistics.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022*

Conference on Empirical Methods in Natural Language Processing, pages 12052–12066, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization.

Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. 2023. Veco 2.0: Cross-lingual language model pre-training with multi-granularity contrastive learning. *arXiv preprint arXiv:2304.08205*.

Kaiyan Zhao, Qiyu Wu, Xin-Qiang Cai, and Yoshimasa Tsuruoka. 2024. Leveraging multi-lingual positive instances in contrastive learning to improve sentence embedding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 976–991, St. Julian's, Malta. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.