Cross-lingual Sentence Embedding using Multi-Task Learning

Koustava Goswami[1]*, **Sourav Dutta**[2], **Haytham Assem**[2],
Theodorus Fransen[1] and **John P. McCrae**[1]

[1] Data Science Institute, National University of Ireland Galway, Ireland
[2] Huawei Research Centre, Dublin, Ireland
{koustava.goswami,theodorus.fransen, john.mccrae}@insight-centre.org
{sourav.dutta2, haytham.assem}@huawei.com

Abstract

Multilingual sentence embeddings capture rich semantic information not only for measuring similarity between texts but also for catering to a broad range of downstream cross-lingual NLP tasks. State-of-the-art multilingual sentence embedding models require large parallel corpora to learn efficiently, which confines the scope of these models. In this paper, we propose a novel *sentence embedding framework* based on an *unsupervised loss function* for generating effective multilingual sentence embeddings, eliminating the need for parallel corpora. We capture semantic similarity and relatedness between sentences using a *multi-task loss function* for training a *dual encoder model* mapping different languages onto the same vector space. We demonstrate the efficacy of an unsupervised as well as a weakly supervised variant of our framework on STS, BUCC and Tatoeba benchmark tasks. The proposed unsupervised sentence embedding framework outperforms even supervised state-of-the-art methods for certain under-resourced languages on the Tatoeba dataset and on a monolingual benchmark. Further, we show enhanced *zero-shot learning* capabilities for more than 30 languages, with the model being trained on only 13 languages. Our model can be extended to a wide range of languages from any language family, as it overcomes the requirement of parallel corpora for training.

1 Introduction

Sentence embeddings provide an efficient way to encode semantic information of text by mapping texts onto a shared vector space, such that sentences with similar meaning are represented by similar representations. With the abundance of data in diverse languages, cross-lingual sentence embedding enable the mapping of multilingual texts into a single unified vector space for a wide range of Natural Language Processing (NLP) tasks. Current sentence embedding methods are predominantly monolingual systems, geared mainly towards English (Conneau et al., 2017; Yin et al., 2020). While there exist multilingual sentence embedding frameworks, they are mostly supervised methods requiring a large parallel corpus for training. For under-resourced languages, there is not sufficient training data to effectively learn a model and we show that our unsupervised approach can better exploit the available unsupervised data, and thus produce better results for under-resourced languages. This is achieved by using a dual-encoder architecture based on word-level semantic similarity score (via Word Mover's Distance) and learning to embed this into a single vector for sentences.

Supervised sentence embedding approaches map parallel sentences from source and target languages into the same vector space by either maximising their cosine similarity or minimising the distance between the generated embeddings (Artetxe and Schwenk, 2019; Reimers and Gurevych, 2020). For example, recent supervised methods using parallel corpus rely on a *teacher-student model* to minimize cross-lingual embedding distance (Reimers and Gurevych, 2020) or an additive margin softmax function based dual sentence encoder to maximally separate the sentences that are true translations from similar overlapping sentences (Yang et al., 2019). Although such methods produced good results, the use of these loss functions in unsupervised settings fails to efficiently capture cross-lingual semantic similarities across sentences. The state-of-the-art unsupervised approach relies on automated machine translation to generate a "pseudo parallel corpus" (Kvapilíková et al., 2020) for training. This method is affected by presence of translation errors and fails to generalize to low-resource languages for which translations are not available.

To alleviate the above challenges, in this paper, we propose *DuEAM*, a cross-lingual sentence embedding framework based on a novel *dual encoder*

*Work started during internship at Huawei Research.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9099–9113
November 7–11, 2021.

architecture with an unsupervised joint loss function using an *anchor-learner* approach, a variant of the teacher-student model. We also depict the performance of a *weakly-supervised variant* of our unsupervised *DuEAM* architecture (obtained by simply changing the training dataset). The weakly-supervised framework to learn semantic relationship between cross-lingual sentences is motivated by the existence of the multilingual natural language inference dataset (XNLI) (Conneau et al., 2018), and the possible creation of such a dataset from existing comparable corpora. The unsupervised *DuEAM* framework learns from randomly chosen sentence pairs from the XNLI dataset (see Sec. 5). Thus, we overcome the need for parallel sentences for multilingual sentence embedding generation. To understand the degree of similarity between monolingual and multilingual sentence pairs during training, the anchor module uses the Word Mover's Distance (WMD) (Kusner et al., 2015) (used as a scalar value during backpropagation), while the learner module is trained to generate sentence embeddings (refer to Fig. 1). Thus, we learn a low-dimensional embedding of the sentences from the more complex encoding generated by WMD. We show that our joint loss formulation effectively captures cross-lingual semantic similarity between sentences by preserving distances between points across languages.

Extensive experiments (in Section 6) on multilingual sentence similarity and parallel sentence mining tasks have showcased the efficacy of our sentence embedding framework. For example, on the cross-lingual STS benchmark (Reimers and Gurevych, 2020), our unsupervised approach achieves state-of-the-art average Spearman rank correlation score of $\mathbf{62.1}$, comparable to the supervised sentence embedding approach of LASER (Artetxe and Schwenk, 2019) with an average of 65.8. In fact, for certain languages, our models are even seen to outperform LASER (e.g., for EN-DE our unsupervised model achieves a Spearman rank correlation score of 64.6 and weakly-supervised model achieves a Spearman rank correlation score of 69.4 compared to 64.2 for LASER). On the BUCC task (bitext mining task) (Zweigenbaum et al., 2017) our model achieves a better F1 score compared to the existing unsupervised model of Kvapilíková et al. (2020). Interestingly, for certain under-resourced languages, we outperform both LASER and multilingual S-BERT (Reimers and Gurevych, 2020) by an average of $\mathbf{10}\%$ on the Tatoeba benchmark. We also show better or comparable performance to LASER even on monolingual classification benchmark tasks. Thus, our model is robust across diverse language families for multilingual sentence embeddings.

In a nutshell, our contributions are: **(i)** *DuEAM*, a novel dual encoder based on an anchor-learner architecture for unsupervised and weakly-supervised multilingual sentence embedding generation, **(ii)** a joint loss function coupling Word Mover's Distance and cosine similarity to capture the degree of text similarity and relatedness between sentence pairs, **(iii)** experimental evaluations, on monolingual as well as several cross-lingual benchmark tasks, depict that our model efficiently captures semantic similarity across languages, and provides state-of-the-art unsupervised performance, comparable with supervised models, **(iv)** robustness in zero-shot transfer learning for low-resource languages across language families, outperforming state-of-the-art supervised approaches on sentence matching tasks in certain scenarios.

2 Related Work

Paragraph vectors were first proposed as sentence embeddings for computing document similarity (Le and Mikolov, 2014). The majority of the current multilingual sentence embedding methods are supervised approaches. There exist some unsupervised sentence embedding frameworks (Zhang et al., 2020; Pagliardini et al., 2018), but are mostly for English sentence embeddings. Initial methods generated sentence embeddings based on neural machine translation system with a shared encoder (Schwenk, 2018; España-Bonet et al., 2017; Schwenk and Douze, 2017). The use of cosine similarities between source and target language parallel sentences was studied by Guo et al. (2018) using a bidirectional dual encoder architecture. Chidambaram et al. (2019) proposed Multilingual Universal Sentence Encoder (mUSE), a dual-encoder model trained on large web-mined translation parallel corpora, along with data from Reddit, Wikipedia, and Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) to learn more context, supporting 16 languages. A translation ranking task was used to identify a correct translation pair, and the architecture assumes 5 hard negative pairs for each sample while training. Subsequently, the LASER (Artetxe and Schwenk,

2019; Schwenk et al., 2019) framework considered a sequence-to-sequence architecture using LSTM networks, and was trained on parallel corpora designed for neural machine translation across 93 languages. Expanding beyond translation-based approaches, the multilingual sentence encoder model of Yang et al. (2020a) was trained for semantic retrieval on three different tasks: multi-feature question-answer prediction, translation ranking, and natural language inference (NLI). Recently, Yang et al. (2020b) proposed Conditional Masked Language Modeling (CMLM) to generate sentence embeddings, by co-training the system with bi-text retrieval and Natural Language Inference (NLI) tasks. To generate sentence embeddings beyond the naïve CLS token and simple pooling strategies of language models, sentence transformer architectures were proposed (Reimers and Gurevych, 2019). For multilingual S-BERT models, Reimers and Gurevych (2020) utilized the teacher-student model where the student model is tuned with a parallel corpus from 50 languages based on knowledge transfer from a fine-tuned teacher model developed by Reimers and Gurevych (2019). LaBSE (Feng et al., 2020) was designed on the BERT architecture and trained on 6 billion sentence pairs by use of additive margin softmax loss with in-batch negative sampling.

Another thread of research for sentence embedding involves improving the alignment of contextual embeddings into a shared vector space using iterative self-supervised learning or tuning with synthetic parallel corpora. Hirota et al. (2020) introduced the Enhancing Multilingual Sentence Embeddings (EMU) framework which tries to semantically enhance pre-trained multilingual sentence embeddings. That is, instead of building sentence embeddings from scratch, EMU fine-tunes pre-trained multilingual sentence embeddings with two major components: enhancement of semantic similarity and multilinguality of sentence embeddings using multlingual adverserial training. Further, Cao et al. (2020) used a parallel corpus as an anchor to align representations in a multilingual language model whereas Wang et al. (2019) used iterative self-learning to perform the task.

Recently, Kvapilíková et al. (2020) proposed an unsupervised method for improving pre-trained cross-lingual context vectors using synthetic parallel sentences and extracted sentence embeddings via mean pooling. However, use of machine translation to generate synthetic parallel data for such methods fails to generalize to low-resourced languages for which translations might be erroneous.

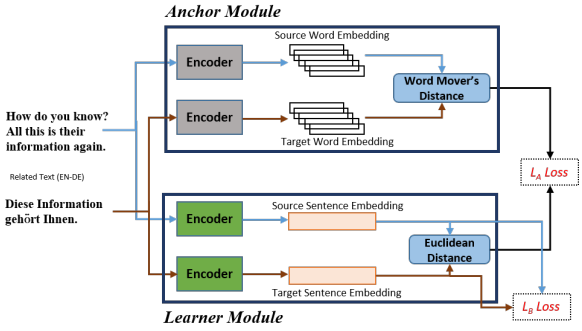

Figure 1: *DuEAM*: Proposed Dual Encoder based Anchor-Learner model with multi-task learning. For training, in our anchor module as well as our learner model, we use the XLM-RoBERTa-base (XLM-R-base) language model (Conneau et al., 2020).

3 *DuEAM* Model

In this section, we describe the components and working of the proposed *Dual Encoder with Anchor Model* (DuEAM) architecture for multilingual sentence embeddings, trained using an unsupervised multi-task joint loss function.

3.1 Dual Encoder with Anchor-Learner

Figure 1 depicts the *dual-encoder* based anchor module, where the same sentence pair is fed as inputs into both the anchor and the learner components. Note, these sentence pairs are not considered to be parallel translations and can even be from either the same language or different languages. Such architectures are well suited to capture semantically and contextually similar sentences and map them close to each other in a shared vector space.

We use word-level semantic knowledge from pre-trained multilingual language models in the anchor module to gauge the semantic similarity between the source (s_i) and target (t_i) sentences. Subsequently, embeddings for s_i and t_i are generated by the learner such that their vector space distances reflect their degree of semantic relatedness. Inspired by MoverScore measure (Zhao et al., 2019), using the pre-trained multilingual word-embeddings, the anchor module computes the semantic similarity between source and target sentences by use of *Word Mover's Distance* (WMD) (Kusner et al., 2015).

The learner module is then trained to generate source and target sentence embeddings such that

their Euclidean distance closely approximates the WMD obtained from the anchor. We force the system to consider the knowledge of our anchor system about the semantic relationships of the sentences at the word-level, which also helps to stabilize the training process as the pre-trained anchor model is fixed and the WMD score is considered as a scalar. Thus, during training, we generate embeddings to *minimize the semantic loss*, $\mathcal{L}_{\mathcal{A}}$ as:

$$\mathcal{L}_{\mathcal{A}} = \frac{1}{N}\sum_{i=1}^{N} \exp^{|\exp^{-d_{euc}(s'_i,t'_i)} - exp^{-d_{wmd}(s_i,t_i)}|} \quad (1)$$

where $d_{wmd}(s_i, t_i) = WMD(s_i, t_i)$ is the Word Mover's Distance between the input source and target sentences s_i and t_i, while $d_{euc}(s'_i, t'_i) = \sqrt{\sum_j \left(s'_{i_j} - t'_{i_j}\right)^2}$ is the Euclidean distance between the generated embeddings s'_i and t'_i (by the learner) corresponding to the source and target sentences respectively. The use of our WMD based semantic loss factor enables *DuEAM* to capture a more compact representation between the sentence embeddings, better capturing cross-lingual semantic relationships in the shared vector space.

3.2 Dual Encoder with Translation Mining

While mapping semantically similar sentences close to each other in the shared vector space, an effective embedding framework should also address the translation ranking problem to efficiently map correct translations of source-target sentence pairs within a compact zone of the vector space. Yang et al. (2019) addressed this problem by introducing hard negative sentence pairs along with parallel data during the training process. Since *DuEAM* is an unsupervised approach, we introduce *translation mining based loss*, $\mathcal{L}_{\mathcal{B}}$, using the cosine similarity score between source and target sentences, to handle translation mining, as:

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{N}\sum_{i=1}^{N} cossim(s'_i, t'_i) \quad (2)$$

3.3 Multi-Task Dual Encoder Learning

To bring both loss functions under the same umbrella, we construct a *multi-task learning setup* where we minimize Eq. 1 while maximizing Eq. 2. Hence, to efficiently generate sentence embeddings, the final multi-task loss function for training *DuEAM* is given by: $\boxed{minimize\ \mathcal{L} = \mathcal{L}_{\mathcal{A}} - \lambda\mathcal{L}_{\mathcal{B}}}$, where λ is the weight parameter.

This multi-task joint learning enables *DuEAM* to effectively capture both cross-lingual semantic similarity and text translation ranking relationship.

Overall, our *unsupervised loss function* aims to learn sentence embeddings such that the Euclidean distance between them are proportional to the semantic distance obtained from WMD, thereby providing a low-dimensional embedding from the more complex word-level similarity space (using Eq. (1)) Additionally, Eq. (2) enables our framework to align sentence embeddings in the cosine space, for translation understanding.

4 Intuitions behind Loss Function

In *DuEAM*, WMD is computed between the contextual token embeddings (of the pair of input sentences) obtained from anchor encoders, without any stopwords removal (as standard while using language models (Conneau et al., 2020)). While multi-lingual language models capture different languages in a common space, WMD captures the contextualized semantic distance between the multi-lingual input sentences. Our use of WMD is motivated by MoverScore (Zhao et al., 2019). Specifically, the use of WMD and cosine in the loss function of *DuEAM* is based on the following intuitions:

- The learner module is trained to generate sentence embeddings such that the Euclidian distance between the learnt sentence embeddings closely approximates the WMD (calculated using token embeddings by anchor) between sentence pairs. This tends to preserve the "relative semantic distance" between the input sentences, enabling *DuEAM* to capture the "semantic relation at word level" within the sentence embeddings obtained.

- Existing methods using parallel sentences for training effectively teach the architecture to learn similar embeddings for similar context – however, the distance in the embedding space between dissimilar sentences are not considered. By using WMD, *DuEAM* generates closer representations for similar sentences, while at the same time forcing dissimilar sentences to have embeddings that are apart in the embedding space. This provides better semantic understanding for improved performance in downstream tasks, as observed in our experiments. For the example, the WMD

between the German sentence "Sie ist keine Lehrerin" ('She is not a teacher') and the English sentence "She is a teacher" is more than that between the German sentence "Sie ist eine Lehrerin" and the English sentence that is a direct translation, "She is a teacher". Thus, the embeddings are different, and *DuEAM* is able to capture negation and other semantic information for better performance.

- The cosine loss enables *DuEAM* to align the learnt sentence embeddings in the cosine space, based on the cosine similarity between the source and target train sentence pairs, to address the translation ranking problem. The weight parameter λ in the final multi-task loss function further controls its effect.

5 Training Dataset

Following Chidambaram et al. (2019), we train our *DuEAM* architecture on the natural language inference dataset – using only the XNLI dataset (Conneau et al., 2018) on 13 languages, without any parallel corpora [1]. We do not consider the entailment-contradiction labels in XNLI during training.

Unsupervised Data. To create the training dataset for unsupervised training of *DuEAM*, we randomly pick sentences from the premises and hypothesis of XNLI dataset and form the input sentence pairs. Hence, this random shuffling along with the absence of sentence pair labels provides no supervision during the training procedure.

Weakly Supervised Data. This training data contains both monolingual and cross-lingual sentence pairs, where the monolingual sentence pairs are same as those of XNLI (without annotated labels). To create cross-lingual sentence pairs, we keep the premises from the source language and replace the hypothesis with target language hypothesis sentences, and vice-versa (example shown in Table 9). Note that this dataset does not contain any parallel cross-lingual sentences, but contains semantically related monolingual and cross-lingual sentences – providing weak supervision.

Validation (using accuracy of finding parallel sentences) is done on held out $1K$ parallel sentences (across languages pairs) from the TED2020 corpus. The *DuEAM* models trained on the above unsupervised and weakly supervised datasets are henceforth denoted as $\text{DuEAM}_{\text{unsupv}}$ and $\text{DuEAM}_{\text{wklysupv}}$ respectively. More details on dataset and training are given in the appendix.

[1] Trained on EN, BG, DE, EL, ES, FR, HI, RU, TH, TR, UR, VI, ZH

Approaches / Languages	EN-EN	ES-ES	EN-DE	EN-TR	EN-ES	EN-FR
Supervised Methods						
XLM-R ← SBERT-nli-stsb	82.5	83.5	78.9	74.0	79.7	78.5
XLM-R ← SBERT-paraphrases	**88.8**	86.3	**84.0**	**80.9**	**83.1**	**84.9**
LASER	77.6	79.7	64.2	72.0	57.9	69.1
LaBSE	79.4	80.8	73.8	72.0	65.5	77.0
mUSE	86.4	**86.9**	82.1	75.5	79.6	82.6
Unsupervised Methods						
mBERT mean	54.4	56.7	33.9	16.0	21.5	33.0
XLM-R mean	50.7	51.8	21.3	9.2	10.9	16.6
Proposed methods						
$\textit{DuEAM}_{\textit{wklysupv}}$	**81.9**	**83.1**	**69.4**	**68.6**	**64.6**	**69.6**
$\textit{DuEAM}_{\textit{unsupv}}$	*80.2*	*81.5*	*64.6*	*63.7*	*58.2*	*62.1*

Table 1: Spearman rank correlation (ρ) results for Semantic Textual Similarity (STS) datasets. The results are reported as $\rho \times 100$, with baseline performances as reported in (Reimers and Gurevych, 2020).

6 Experimental Evaluation

We evaluate the performance of our proposed *DuEAM* framework on the following 3 benchmark tasks: (a) **STS:** monolingual and cross-lingual *semantic textual similarity*; (b) **BUCC:** *bitext mining* to extract parallel sentences; and (c) **Tatoeba:** cross-lingual *parallel sentence matching*.

We compare our model with the following supervised and unsupervised state-of-the-art approaches: (i) **mBERT / XLM-R** – language model with mean pooling, (ii) **mUSE** – dual-encoder transformer architecture (Chidambaram et al., 2019), (iii) **LASER** – encoder-decoder architecture using LSTM (Artetxe and Schwenk, 2019), (iv) **LaBSE** – dual-encoder model based on BERT (Feng et al., 2020), (v) **XLM-R ← SBERT-nli-stsb / XLM-R ← SBERT-paraphrases** – sentence transformer models (Reimers and Gurevych, 2020).
Further details on training setup and baseline methods can be found in Section A.

6.1 Multilingual Semantic Textual Similarity

Understanding semantic textual similarity between monolingual and cross-lingual datasets is one of the major tasks for a sentence embedding model. We evaluate our model against the **STS** benchmark dataset (Cer et al., 2017), containing sentence pairs with scores indicating how semantically similar the sentences are. The SemEval dataset consists of annotated sentences for EN-EN, AR-AR, ES-ES, EN-AR, EN-ES, and EN-TR language pairs.

Further, we also use the EN-DE, EN-IT and EN-NL test sets from multilingual SBERT (Reimers and Gurevych, 2020). For evaluation, we compute the cosine similarity between sentence pair embeddings and obtain the *Spearman rank correlation*, ρ across the computed similarities and gold scores.

As shown in Table 1, the unsupervised baselines based on the language models (mBERT and XLM-R) perform quite poorly, suggesting that the obtained cross-lingual sentence embeddings are not well aligned in the vector space. While trained with multi-task learning, *DuEAM* achieved a significant improvement both for monolingual and cross-lingual datasets. For cross-lingual settings, both the *DuEAM* models significantly outperform the unsupervised models, with an average improvement of $\mathbf{41.9}$ and $\mathbf{37.2}$ respectively on Spearman rank correlation (ρ) score. Similarly, on the *monolingual* datasets (EN-EN and ES-ES), our models achieve an improvement of $\mathbf{26.9}$ and $\mathbf{21.4}$ points (on average), based on the rank correlation score.

It is interesting to note that *DuEAM* achieves better results, compared to the supervised LASER and LaBSE approaches, for both the monolingual datasets (EN-EN and ES-ES). In cross-lingual settings, in certain cases, both $\text{DuEAM}_{\text{unsupv}}$ and $\text{DuEAM}_{\text{wklysupv}}$ are seen to outperform LASER (e.g., EN-DE and EN-ES language pairs).

6.1.1 Zero Shot Testing on STS benchmark

An important property of the embedding techniques is *"zero shot learning"*, i.e., to robustly generalize to languages that the model has not been trained on, by inherent knowledge transfer from the other languages. To study the efficiency of *DuEAM* in zero-shot scenarios, in this setting, we train the weakly-supervised model only on the EN-DE (English-German) training dataset (as in Section 5) and test on the other language pairs, including monolingual sentence similarity for ES-ES and AR-AR.

From Table 2, we can see that even for zero shot learning *DuEAM* outperforms the unsupervised baseline models, across all the monolingual and cross-lingual STS datasets, with improvements in Spearman rank correlation score of around $\mathbf{20}$. Thus, our architecture based on dual encoder with multi-task learning provides better cross-lingual sentence embeddings, making it robust across diverse languages with improved performance.

Models	ES-ES	AR-AR	EN-ES	EN-AR	EN-TR	EN-FR	EN-NL	EN-IT
mBERT mean	56.7	50.9	21.5	16.7	16.0	33.0	35.6	34.0
XLM-R mean	51.8	25.7	10.9	17.4	9.2	16.6	26.0	22.9
DuEAM$_{wklysupv}$	**78.64**	**69.67**	**56.54**	**54.29**	**58.35**	**62.03**	**67.61**	**59.8**

Table 2: Zero-shot Spearman rank correlation ρ results for Semantic Textual Similarity (STS) datasets, where models are trained on EN-DE non-parallel data.

Approaches / Languages	DE-EN	FR-EN	RU-EN	ZH-EN
Supervised Methods				
XLM-R ← SBERT-nli-stsb	86.8	84.4	86.3	85.1
LASER	95.4	92.4	92.3	91.7
LaBSE	**95.9**	**92.5**	**92.4**	**93**
mUSE	88.5	86.3	89.1	86.9
Unsupervised Methods				
mBERT mean	44.1	47.2	38	37.4
XLM-R mean	5.2	6.6	22.1	12.4
(Kvapilíková et al., 2020)	80.2	78.8	77.1	67.0
Proposed Methods				
DuEAM$_{wklysupv}$	**84.9**	**81.3**	**82.0**	**78.6**
DuEAM$_{unsupv}$	*80.9*	*79.3*	*78.4*	*70.0*

Table 3: F1 score on BUCC bitext mining task. Baseline results taken from (Reimers and Gurevych, 2020).

6.2 Bitext Mining Task

Efficient multilingual sentence embeddings should have a good understanding of sentence parallelism and should be able to retrieve good translation pairs across corpora in different languages. Intuitively, sentence translation pairs should be equivalent in terms of semantic similarity, and hence their cross-lingual embeddings should be very similar.

To evaluate the performance of our method, we conduct experiments on the **BUCC** benchmark mining task – parallel sentence extraction from two different monolingual corpora. We use the data available from the 2018 shared task, consisting of corpora for four language pairs (FR-EN, DE-EN, RU-EN, and ZH-EN), with a subset of parallel sentences demarked as the gold mapping for each language pair. The data is split into train and test set, and the training data is used to find a threshold for the scoring function, such that sentence pairs above the threshold are returned as parallel sentences. Performance is measured using *F1 score*. Similar to Reimers and Gurevych (2020), in this setting, we use the margin scoring function as:

$$\text{score}(x, y) = \text{margin}(\cos(x, y), \cos^*(x, y)), \text{ with}$$

$$\cos^*(x, y) = \sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2K} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2K}$$

where *x, y* are the sentence embeddings, $\text{NN}_k(x)$ is the *k* nearest neighbours of *x* in other languages

excluding duplicates, and margin$(a, b) = \frac{a}{b}$.

Table 3 shows the performance of the approaches on the BUCC task. For the unsupervised setting, we observe that for all language pairs *DuEAM* performs significantly better than XLM-R mean and mBERT mean methods, with an improvement of nearly **35.6** F1 score over mBERT. Additionally, we compare our model with the recent approach by Kvapilíková et al. (2020), specifically trained for bitext mining task with synthetic parallel data. We see that *DuEAM* also outperforms these unsupervised models across all language pairs.

Further, our weakly-supervised model achieves competitive results compared to the supervised model of XLM-R ← SBERT-nli-stsb, trained with large parallel datasets. Observe that LASER and LaBSE achieve high accuracy, as they are specifically designed and trained to identify translations between languages. On the other hand, although *DuEAM* is not trained with any parallel data, we are able to effectively extract parallel sentences (bitext mining task) owing to our multi-task learning.

6.3 Cross-lingual Parallel Sentence Matching

In this section, we compare the performance of the approaches in extracting parallel sentences using the **Tatoeba** benchmark of Artetxe and Schwenk (2019). From Table 4, on well-resourced languages, we observe *DuEAM* to perform significantly better than the unsupervised approach of Kvapilíková et al. (2020), with results comparable to the supervised methods of SBERT and LASER – similar as above – efficiently extracting parallel sentences.

6.3.1 Tatoeba Under-Resourced Languages

We now evaluate the robustness of the approaches for extracting parallel sentences for under-resourced languages on the Tatoeba benchmark. In this setting, we consider two scenarios – (i) *zero-shot learning* and (ii) training on *small scale* non-parallel datasets. We compare the results with the supervised models of SBERT (Reimers and Gurevych, 2020) and LASER (Artetxe and Schwenk, 2019) respectively, with different baseline languages for LASER for which it has been pre-trained and results have been published.

Zero-shot Transfer. We perform zero-shot transfer on different under-resourced languages: (i) Telugu (TE, Dravidian family), (ii) Tagalog (TL, Malayo-Polynesian family), (iii) Irish (Gaelic) (GA, Celtic family), and (iv) Afrikaans (AF, Germanic family). We observe from Table 5(a) that

Model	DE	HI	ZH	EL
XLM-R ← SBERT-paraphrases	98.7	96.4	95.0	95.5
LASER	99.0	94.7	95.4	95.0
Kvapilíková et al. (2020)	83.1	53.4	-	51.3
DuEAM$_{wklysupv}$	96.0	92.9	90.2	87.4
DuEAM$_{unsupv}$	93.4	83.5	85.2	82.0

Table 4: Average accuracy on Tatoeba dataset in both directions (EN to target language and vice-versa). Here ZH refers to Mandarin Chinese. Baseline results taken from (Reimers and Gurevych, 2020).

Model	AF	TE	TL	GA
XLM-R ← SBERT-para	84.2	89.1	32.4	18.6
DuEAM$_{wklysupv}$	**84.8**	**90.6**	**60.6**	**42.0**
DuEAM$_{unsupv}$	79.9	78.6	*56.8*	*35.0*

(a)

Model	KA	AM
LASER	35.9	42.0
DuEAM$_{wklysupv}$	**76.4**	**56.0**
DuEAM$_{unsupv}$	*70.7*	*46.4*

(b)

Table 5: Average accuracy on Tatoeba data in both directions (EN to target language and vice versa) for (a) zero-shot learning, and (b) small training set. Baseline results as in (Reimers and Gurevych, 2020).

for Tagalog and Irish (Gaelic) both *DuEAM* models performed significantly better than the supervised multilingual S-BERT with an improvement of around 25% on average for weakly-supervised model and 21% on average for unsupervised model, while for Afrikaans and Telugu, we achieved 1.5% better accuracy on average for weakly-supervised model. These results show that, even without explicit learning for different under-resourced languages, our models can robustly handle zero-shot learning across different language families for generating efficient sentence embeddings.

Small Scale Dataset Training. To explore the model performance while trained on small datasets (for scenarios where limited data is available for under-resourced languages), we experiment on two under-resourced languages: Georgian (KA, Kartvelian family) and Amharic (AM, Ethio-Semitic family). We train *DuEAM* with *only* $20K$ *non-parallel* EN-KA and EN-AM sentence pairs, whereas the baseline supervised model LASER is trained with $296K$ and $88K$ parallel sentence pairs respectively. In Table 5(b), we see that our models outperform LASER for both Georgian and Amharic, achieving higher accuracy on under-resourced languages even when trained on a much smaller non-parallel dataset. In fact, weakly-supervised *DuEAM* performs better than SBERT (Reimers and Gurevych, 2020), producing an accuracy of 72.4% on KA. The anchor-learner architecture with the unsupervised joint loss function provides such robustness and better cross-lingual understanding in *DuEAM*. Extensive results on 58

Model	MR	SUBJ	TREC	SST2
XLM-R ← SBERT-paraphrases	**81.26**	**93.89**	91.2	**87.7**
LASER	75.29	92.07	91.0	79.9
DuEAM$_{unsupv}$	76.28	92.86	**92.2**	81.1

Table 6: Evaluation accuracy on a subset of SentEval benchmark (results based on 10-fold cross-validation).

languages of Tatoeba benchmark for trained and zero-shot scenarios can be found in the appendix.

6.4 Monolingual Classification Performance

A multilingual sentence embedding framework is expected to produce efficient results in monolingual settings. To evaluate the performance on monolingual classification tasks, we now study the performance of the unsupervised variant of *DuEAM* on the SentEval benchmark (Conneau and Kiela, 2018). We compare *DuEAM* to other sentence embedding frameworks on four tasks : (i) **MR**: Movie reviews positive/negative sentiment analysis (Pang and Lee, 2005), (ii) **SUBJ**: Subjectivity/objectivity prediction of reviews (Pang and Lee, 2004), (iii) **TREC**: Question type classification on six classes (Li and Roth, 2002), and (iv) **SST2**: Stanford binary sentiment classification (Socher et al., 2013; Reimers and Gurevych, 2019).

In Table 6, we can see quite satisfactory result produced by *DuEAM* framework. On every task, the *DuEAM*$_{unsupv}$ model surpasses the performance of the supervised LASER model whereas in case of the **TREC** task the results are better than even the supervised multilingual S-BERT model.

Overall, the monolingual and multi-lingual experimental results depict *DuEAM* to effectively capture cross-lingual semantic understanding (without parallel training data) to generate efficient sentence embeddings by alignment of multiple languages in the same vector space. Observe that *DuEAM* is trained on only 1GB of data, while other supervised techniques are trained on around 10x or more data.

7 Ablation Study

We now study the effects of different components of *DuEAM* on the quality of generated embeddings.

Necessity of Multi-task Learning. One of the important features of *DuEAM* is *multi-task joint learning* via the dual-encoder based anchor-learner architecture. To explore the necessity of the different factors for our learning loss function, we use the Tatoeba dataset for DE-EN, FR-EN, and HI-EN language pairs. We train weakly-supervised

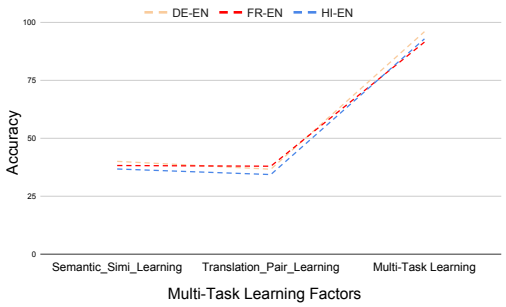

Figure 2: Average accuracy on Tatoeba across language pairs with individual objectives of mutli-task learning.

Zero-Shot	1K	5K	10K	15K	20K	25K	30K
71.8	72.9	73.5	74.8	75.9	76.41	76.40	76.42

Table 7: Average accuracy of *DuEAM*$_{wklysupv}$ on Tatoeba (in both directions) for Georgian language (KA) with varying training data sizes.

DuEAM$_{wklysupv}$ in three variants: (i) with only the anchor module loss term $\mathcal{L}_{\mathcal{A}}$ (Eq. 1), which considers semantic similarity between sentences, (ii) using only the translation mining loss term $\mathcal{L}_{\mathcal{B}}$ (Eq. 2), which identifies the best translation pairs, and (iii) the full multi-task learning objective $\mathcal{L}$.

From Figure 2, we observe that the learning objective factors $\mathcal{L}_{\mathcal{A}}$ and $\mathcal{L}_{\mathcal{B}}$ individually perform quite poorly. However, the proposed multi-task training performs efficiently providing a high accuracy in the range of 92% to 96%. Similar results are observed across the language pairs considered.

Training Dataset Size. Table 5(b) depicts that *DuEAM* outperforms LASER on under-resourced languages with minimal training. To understand the impact of the size of the training dataset, we evaluated the performance of *DuEAM* on Tatoeba data for Georgian (KA), with varying training sizes.

In Table 7 we see a healthy performance improvement when our weakly-supervised model is trained with the language-specific dataset, with around 4.5% improvement over zero-shot learning given training data of size $20K$ sentences. Thereafter, the improvement is seen to be incremental. Thus, although *DuEAM* demonstrates zero-shot learning capabilities, a small amount of language-specific data further boosts the performance.

Training Dataset Type. We explore the performance of the unsupervised loss function in *DuEAM* on various training data scenarios. We consider EN, DE, FR, and ES languages under 3 training settings: (i) $25K$ sentences from XNLI (weakly supervised),

Dataset	DE	FR	ES
$25K$ XNLI dataset	88.7	82.6	86.5
$25K$ TED2020 parallel dataset	89.2	83.8	86.9
$12.5K$ XNLI + 12.5K parallel dataset	90.2	85.1	89.5

Table 8: Average accuracy of $DuEAM_{wklysupv}$ on Tatoeba (considering both directions).

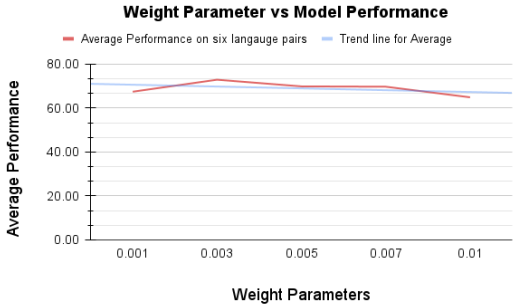

Figure 3: Average Spearman rank correlation (ρ) results for Semantic Textual Similarity (STS) datasets for six language pairs.

(ii) $25K$ parallel sentences from TED2020 (supervised), and (iii) $12.5K$ sentences from each of the datasets. Table 8 depicts similar performances when trained on XNLI or on a parallel corpus alone, while a combination outperforms the others – showcasing the stability of our proposed *unsupervised loss joint loss function* (on training data), removing the dependency on parallel training datasets.

Weight Parameter Value Selection. Weight parameter selection while training the multi-task model is very important. We experimented with the weight parameter over a range of values and set to the value for which the model has performed best while training. We have given a snapshot of the model performances over the different weight parameters in Figure 3. We have calculated the average Spearman rank correlation (ρ) results for STS datasets for six language pairs. From the figure we can see that the best model performance achieved with weight parameter 0.003. Higher weight parameter value decreased the performance. This helps us to understand the importance of the weight parameter while training the model in multi-task settings.

8 Discussion: Semantic Similarity

In general, applications use cosine similarity between sentence embeddings to gauge the semantic textual similarity. We provide a performance analysis of *DuEAM* based on the raw cosine similarity score on Tatoeba DE-EN data. For example, the German sentence "das ist der Geburtstag von Muiriel!" ("That is the birthday of Muiriel") has the highest cosine similarity with its English translation "it is Muiriel's birthday!", although the sentence "Happy birthday, Muiriel!" is very similar (refer Table 10 in appendix). This depicts that *DuEAM* can capture fine-grained semantic difference among similar sentences.

On the other hand, for the German sentence "Das Wesen der Freiheit liegt in der Mathematik." we obtain a higher cosine similarity score for the English sentence "The essence of mathematics is liberty.". In fact, the true translation "The essence of freedom lies in mathematics" (achieving the highest cosine-similarity but absent in the Tatoeba dataset) is closer to "The essence of liberty is mathematics.". Although the similarity score is almost equal, our model is unable to identify the correct word ordering in highly overlapping sentences as the WMD measure is inherently word-order agnostic. Multilingual SBERT too fails in this scenario, but with a higher difference in cosine-similarity between sentences, 0.01 compared to 0.001 in *DuEAM* (using the translation mining $\mathcal{L}_{\mathcal{B}}$ loss factor). Use of Wikipedia dumps for training such sentence embedding models forms an interesting future study.

9 Conclusion

This paper proposed an *unsupervised loss function based DuEAM framework* for *multilingual sentence embeddings* based on dual encoder with anchor-learner model via multi-task learning. Experiments on monolingual and cross-lingual benchmarks showcase the efficacy of our sentence embeddings in capturing semantic similarities across languages. We demonstrate that *DuEAM* significantly outperforms existing unsupervised models for textual similarity understanding. We also depicts robustness in *zero-shot learning* and *limited training*, for catering to under-resourced languages, and achieve results better or comparable to existing supervised methods in certain cases.

Acknowledgements

This publication was supported by a research grant from Irish Research Council Grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages) and Science Foundation Ireland (SFI) under Grant SFI/12/RC/2289_P2 (Insight_2).

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *ICLR*.

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 *SemEval-2017*, pages 1–14.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *RepL4NLP@ACL 2019*, pages 250–259.

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 *ACL*, pages 8440–8451.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of NMT-derived interlingual embeddings and their use in parallel sentence identification. *IEEE J. Sel. Top. Signal Process.*, 11:1340–1350.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

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.

Wataru Hirota, Yoshihiko Suhara, Behzad Golshan, and Wang-Chiew Tan. 2020. Emu: Enhancing multilingual sentence embeddings with semantic specialization. In *AAAI*, pages 7935–7943.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondrej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2020, Online, July 5-10, 2020*, pages 255–262. Association for Computational Linguistics.

Quoc V. Le and Tomás Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 528–540.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 228–234. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 157–167. Association for Computational Linguistics.

Holger Schwenk, Douwe Kiela, and Matthijs Douze. 2019. Analysis of joint multilingual sentence representations and semantic k-nearest neighbor graphs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6982–6990. AAAI Press.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Shuai Wang, Lei Hou, Juanzi Li, Meihan Tong, and Jiabo Jiang. 2019. Learning multilingual sentence embeddings from monolingual corpus. In *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 346–357. Springer.

Yinfei Yang, Gustavo Hernández Ábrego, 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 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020a. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 87–94. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2020b. Universal sentence representation learning with conditional masked language model. *CoRR*, abs/2012.14388.

Xiaoya Yin, Wu Zhang, Wenhao Zhu, Shuang Liu, and Tengjun Yao. 2020. Improving sentence representations via component focusing. *Applied Sciences*, 10(3):958.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1601–1610.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578.

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.

Appendix

A Training setup

For training, in our anchor module as well as our encoder model, we use the XLM-RoBERTa (XLM-R) language model, with its publicly available $250K$ shared vocabulary. *DuEAM* is trained for 5 epochs with a batch size of 64, $10K$ warm-up steps, and a learning rate of $2e-5$. We set λ (parameter in our joint loss function) to 0.003, and the training was performed on a 24GB Titan RTX GPU for about 12 days. Finally, we apply *MEAN* pooling on the final layer of the encoder to get the sentence embeddings. To test the model on SentEval toolkit, we have set k-fold to 10 keeping epoch size to 10, batch size to 64 and 0.2 dropout rate.

B Baseline Models

We have compared the performance of *DuEAM* on benchmark datasets with multiple supervised and unsupervised baseline models which are as follows:

- **mBERT / XLM-R mean:** We use publicly available mBERT and XLM-RoBERTa (XLM-R) language models trained on large datasets (no parallel sentences considered during pre-training phase). We consider mean pooling of the output layer as the sentence embedding.
- **mUSE:** Multilingual Universal Sentence Encoder uses a dual-encoder transformer architecture to generate sentence embeddings trained using parallel corpora for 16 languages.
- **LASER:** Language Agnostic Sentence Representation is designed based on an encoder-decoder architecture using LSTM networks. The model is trained with big parallel datasets and performs max-pooling to get the sentence embedding from the stacked network. It supports 93 languages.
- **LaBSE:** Language-agnostic BERT Sentence Embedding (LaBSE) is a dual-encoder model based on BERT. The model was trained on 6 billion parallel sentence pairs over 109 languages.
- **XLM-R ← SBERT-nli-stsb / XLM-R ← SBERT-paraphrases:** The sentence transformer models generate sentence embeddings using the *teacher-student* architecture, where the XLM-R student model is trained with parallel sentences for across languages with knowledge transfer from the fine-tuned English SBERT-nli-stsb or SBERT-paraphrases as the teacher model.

Description and details of the above models can be publicly obtained from the links as presented in Table 11.

Premise	Hypothesis	Type
How do you know? All this is their information again.	This information belongs to them.	*Monolingual (EN-EN)*
- woher weißt du das ? All das sind ihre Informationen.	Diese Information gehört Ihnen.	*Monolingual (DE-DE)*
How do you know? All this is their information again.	Diese Information gehört Ihnen.	*Cross-lingual (EN-DE)*
- woher weißt du das ? All das sind ihre Informationen.	This information belongs to them.	*Cross-lingual (DE-EN)*

Table 9: Weakly-supervised training dataset example from XLNI for English-German.

Cos. Simi.	Sentence pairs	Results
0.9961	**DE:** das ist der Geburtstag von Muiriel! **EN:** it is Muiriel's birthday!	**True Positives**
0.9465	**DE:** das ist der Geburtstag von Muiriel! **EN:** Happy birthday, Muiriel!	
0.9870	**DE:** Das Wesen der Freiheit leigt in der Mathematik. **EN:** The essence of Mathematics is liberty.	**False Positives**
0.9860 *(0.9971)*	**DE:** Das Wesen der Freiheit leigt in der Mathematik. **EN:** The essence of liberty is mathematics. ***(Correct EN Translation: The essence of freedom lies in Mathematics.***	

Table 10: Raw cosine similarity value on the Tatoeba DE-EN dataset.

C Training Data

To create the weakly-supervised training dataset, we keep monolingual sentence pairs same as those of XNLI dataset. To create cross-lingual sentence pairs, we keep premises from the source language and replace the hypothesis with target language hypothesis sentences, and vice-versa. In the example in Table 9, for language pair EN-DE, the premise is taken from English while the hypothesis is from German and vice-versa. We do not consider any labels to train our model.

D Language Codes

We have empirically evaluated existing sentence embedding techniques with the proposed *DuEAM* architecture on several languages across diverse language families, including low-resourced languages. We reported results for 8 language pairs on the STS benchmark and for 4 language pairs on the BUCC benchmark. On the Tatoeba dataset, we conducted experiments for the full set of 58 languages under different use-case scenarios. Tables 12 and 13 list the languages along with their codes as provided in the benchmark datasets and as presented in the main body of our paper.

Datasets	Link
mBERT	https://huggingface.co/bert-base-multilingual-cased
XML-R	https://huggingface.co/transformers/model_doc/xlmroberta.html
LASER	https://tfhub.dev/google/LaBSE/1
XLM-R ← SBERT-nli-stsb	https://www.sbert.net/docs/pretrained_models.html
XLM-R ← SBERT-paraphrases	https://www.sbert.net/docs/pretrained_models.html
mUSE	https://tfhub.dev/universal-sentence-encoder-xling/many

Table 11: Source of the competing approaches and models used as baselines.

E Tatoeba Results

We have performed all our experiments on the 58 languages of the Tatoeba test datasets. Evaluation for the parallel sentence matching task is done by finding the most similar sentence between two languages based on their cosine similarity. We have calculated accuracy in both directions (English to target language and vice versa), and have reported the average accuracy of the two. We have reported performance results of the different approaches based on three settings:
(i) model performance on languages that it has been trained on,
(ii) zero-shot model performance on untrained languages, while compared with supervised trained models, and
(iii) model performance on under-resourced languages for which it is untrained.

In the main body of the paper, we reported snapshots of the results obtained across a few of the languages (taken across varied language families). Here we report the full evaluation results across all the 58 languages. Baseline supervised models are taken from (Reimers and Gurevych, 2020).

Performance on Trained Languages: Table 14 reports the performance of the models across 12 languages. We have compared our model with supervised baseline and unsupervised baseline models. We can see that across all 12 languages our model achieved high accuracy compared to unsupervised model. Our model also achieved comparative results with supervised model XLM-R ← SBERT-nli-stsb and LASER for some languages.

Zero-shot Transfer: We have compared our model accuracy on 30 untrained languages and compared with baseline models. Table 15 shows that for all 30 languages our model has achieved state-of-the-art unsupervised results. While the supervised models are trained on the parallel datasets for these languages, for some languages our model achieved comparative results even in zero-shot settings.

Zero-shot Transfer on Under-Resourced Languages: While our model has achieved high accuracy on wide range of languages, we have compared our model with supervised baseline on 16 under-resourced languages from different language families for zero-shot transfer. From Table 16 we can observe that across all languages weakly-supervised *DuEAM* achieved higher accuracy than the supervised baseline of XLM-R ← SBERT-paraphrases. In fact, our unsupervised *DuEAM* performed better than XLM-R ← SBERT-paraphrases for most of the languages.

Overall, our unsupervised and weakly-supervised *DuEAM* perform significantly better than the existing unsupervised approach, and is comparable with the supervised models across diverse languages. Our model also efficiently supports zero-shot transfer learning and is robust for under-resourced languages.

Language Code	Language	Language Code	Language	Language Code	Language	Language Code	Language
EN	English	DE	German	FR	French	HI	Hindi
AR	Arabic	ES	Spanish	TR	Turkish	NL	Dutch
IT	Italian	RU	Russian	ZH	Chinese		

Table 12: List of languages and their codes as provided in the STS and BUCC dataset.

Language Code	Language	Language Code	Language	Language Code	Language	Language Code	Language
bul	Bulgarian	dan	Danish	mkd	Macedonian	epo	Esperanto
cmn	Mandarin Chinese	est	Estonian	mon	Mongolian	eus	Basque
deu	German	fin	Finnish	nob	Norwegian Bokmål	gla	Scottish Gaelic
ell	Greek	glg	Galician	pes	Persian	isl	Icelandic
fra	French	heb	Hebrew	por	Portuguese	jav	Javanese
hin	Hindi	hrv	Croatian	ron	Romanian	khm	Khmer
rus	Russian	hun	Hungarian	slk	Slovak	lat	Latin
spa	Spanish	hye	Armenian	slv	Slovenian	swg	Swabian
tur	Turkish	ita	Italian	sqi	Albanian	swh	Swahili
tha	Thai	jpn	Japanese	srp	Serbian	uzb	Uzbek
urd	Urdu	kat	Georgian	pol	Polish	war	Waray
vie	Vietnamese	kor	Korean	ind	Indonesian	xho	Xhosa
ara	Arabic	lit	Lithuanian	bre	Breton	yid	Yiddish
cat	Catalan	lvs	Latvian	ceb	Cebuano		
ces	Czech	mar	Marathi	cym	Welsh		

Table 13: List of languages and their codes as provided in the Tatoeba dataset.

Model / Languages	bul	cmn	deu	ell	fra	hin	rus	spa	tur	tha	urd	vie
Supervised Approaches												
XLM-R ← SBERT-paraphrases	94.0	95.0	98.7	**95.5**	94.7	**96.4**	93.5	**98.0**	97.2	**96.3**	**92.2**	**97.2**
LASER	**95.0**	**95.4**	**99.0**	95.0	**95.6**	94.7	**94.6**	**98.0**	**97.5**	95.4	81.9	96.8
Unsupervised Approaches												
Kvapilíková et al. (2020)	56.0	–	83.1	51.3	–	53.4	–	–	–	–	43.7	–
Proposed Approaches												
DuEAM$_{wklysupv}$	**86.0**	**90.2**	**96.0**	**87.4**	**91.5**	**92.9**	**90.0**	**93.0**	**89.6**	**90.1**	**77.8**	**92.0**
DuEAM$_{unsupv}$	**82.5**	**85.2**	**93.4**	**82.0**	**87.7**	**83.5**	**85.5**	**89.5**	**84.1**	**82.4**	**67.9**	**89.6**

Table 14: Average accuracy on the Tatoeba test set in both directions (EN to target language and vice versa) on **trained languages**.

Model / Languages	ara	cat	ces	dan	est	fin	glg	heb	hrv	hun	hye
Supervised Approaches											
XLM-R ← SBERT-paraphrases	87.7	**96.4**	96.3	**96.2**	95.8	**96.4**	**96.0**	88.4	97.0	94.7	**91.3**
LASER	**92.0**	95.9	**96.5**	96.0	**96.7**	96.3	95.5	**92.2**	**97.2**	**96.0**	36.1
Unsupervised Approaches											
Kvapilíková et al. (2020)	41.1	66.9	53.5	–	39.0	47.5	66.9	–	68.2	–	–
Proposed Approaches											
DuEAM$_{\text{wklysupv}}$	**70.7**	**83.3**	**85.3**	**92.3**	**73.0**	**88.70**	**85.0**	**73.1**	**88.7**	**86.1**	**79.0**
DuEAM$_{\text{unsupv}}$	**61.6**	**80.0**	**79.80**	**90.0**	**70.0**	**83.7**	**80.7**	**69.0**	**84.3**	**81.5**	**74.1**

Model / Languages	ita	jpn	kat	kor	lit	lvs	mar	mkd	mon	nob	pes
Supervised Approaches											
XLM-R ← SBERT-paraphrases	94.9	90.7	**91.4**	**90.1**	95.8	**96.4**	91.0	92.2	**91.7**	98.0	**94.8**
LASER	**95.3**	**94.2**	35.9	88.9	**96.2**	95.4	**91.5**	**94.7**	8.2	**98.8**	93.4
Unsupervised Approaches											
Kvapilíková et al. (2020)	–	54.4	41.4	–	43.9		37.3	–	29.0	–	–
Proposed Approaches											
DuEAM$_{\text{wklysupv}}$	**85.7**	**84.2**	**71.7**	**81.3**	**83.2**	**81.2**	**78.9**	**75.2**	**74.7**	**94.8**	**88.9**
DuEAM$_{\text{unsupv}}$	**83.1**	**77.4**	**68.2**	**75.8**	**78.9**	**76.6**	**73.4**	**71.2**	**71.5**	**93.2**	**83.4**

Model / Languages	por	ron	slk	slv	sqi	srp	pol	ind
Supervised Approaches								
XLM-R ← SBERT-paraphrases	94.8	96.4	96.2	95.5	97.5	93.8	97.0	94.1
LASER	**95.2**	**97.4**	**96.6**	**95.9**	**98.0**	**95.3**	**97.8**	**94.5**
Unsupervised Approaches								
Kvapilíková et al. (2020)	–	–	–	–	–	–	–	64.9
Proposed Approaches								
DuEAM$_{\text{wklysupv}}$	**91.2**	**88.5**	**86.2**	**80.5**	**79.9**	**83.7**	**90.4**	**89.5**
DuEAM$_{\text{unsupv}}$	**89.5**	**87.0**	**80.2**	**77.2**	**76.6**	**80.3**	**88.4**	**87.7**

Table 15: Average accuracy on the Tatoeba test set in both directions (EN to target language and vice versa) on **untrained languages**. All the baseline models are trained on parallel training datasets.

Model / Languages	bre	ceb	cym	epo	eus	gla	isl	jav	khm	lat	swg
XLM-R ← SBERT-paraphrases	10.1	11.7	34.9	68.8	48.6	7.5	75.8	37.0	64.8	28.0	**33.9**
DuEAM$_{\text{wklysupv}}$	**11.5**	**14.8**	**52.7**	**79.2**	**66.0**	**21.9**	**81.9**	**40.2**	**65.7**	**44.0**	**33.9**
DuEAM$_{\text{unsupv}}$	9.8	**12.3**	**46.0**	**74.0**	**58.2**	**16.0**	**78.5**	36.1	52.7	**38.9**	31.2

Model / Languages	swh	uzb	war	xho	yid
XLM-R ← SBERT-paraphrases	27.6	32.6	11.4	11.6	52.7
DuEAM$_{\text{wklysupv}}$	**40.2**	**40.1**	**13.0**	**15.5**	**53.7**
DuEAM$_{\text{unsupv}}$	**33.3**	**35.5**	10.8	**14.9**	46.7

Table 16: Average accuracy on the Tatoeba test set in both directions (EN to target language and vice versa) on **untrained under-resourced languages**.