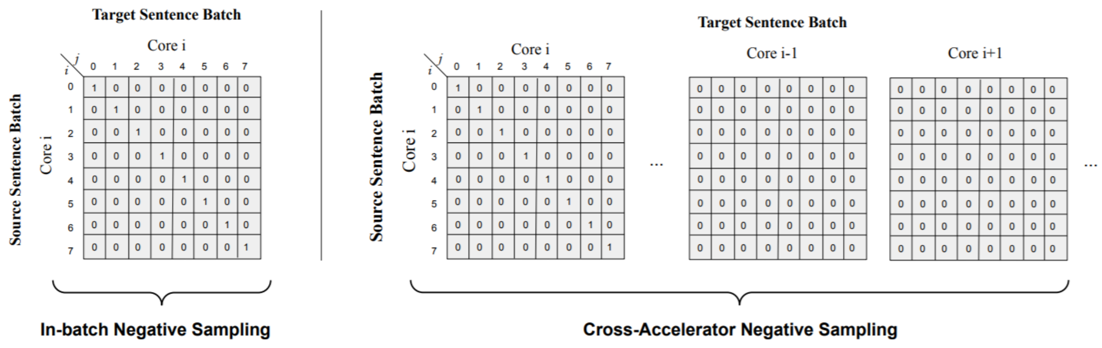
Analysis → (link)

1. They present a novel method for training multilingual sentence-level embeddings combining existing state-of-the-art methods for multilingual sentence embeddings with MLM and translation language model (TLM)
2. Following are their contributions :
    a. To improve the performance of a dual encoder translation ranking model to state-of-the-art performance on bi-text mining, a mix of pre-training and finetuning methods were used.
    b. 109 languages are covered by a single massively multilingual model.Even in zeroshot instances, and demonstrating cross-lingual transfer .
    c. A comprehensive analysis and ablation research was conducted to determine the influence of different data quality, data amount, pre-training, and negative sampling techniques.
3. Similar to Muril , even they have two corpus
    a. Monolingual Corpus(for MLM)
    b. Bilingual Translation Pairs (for TLM)
4. Models
    a. Bidirectional Dual Encoder with Additive Margin Softmax
        i. They use in-batch negative sampling to train bidirectional dual encoders with additive margin softmax loss.
        ii. Loss is asymmetric and is determined by whether the softmax is located over the source or the target.
    b. Cross-Accelerator Negative Sampling
        i. **While data parallelism allows us to use several accelerators to raise the effective batch size, the batch size on a single core has no influence on the batch size across multiple accelerators.
        ii. They introduce **cross-accelerator** negative sampling.
    c. Pre-training and parameter sharing
        i. Masked Language Model (MLM) and Translation Language Model (TLM) are used to train the encoder on monolingual data and bilingual translation pairs, respectively.
        ii. We use a three-stage progressive stacking approach to train a L layer transformer encoder, first learning an L/4 layer model, then L /2 layers, and lastly all L layers.
        iii. The parameters of the models learned in the previous stages are transferred to the models used in the following phases.

Cross Accelerator Architecture



i \ j	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0
3	0	0	0	1	0	0	0	0
4	0	0	0	0	1	0	0	0
5	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	1

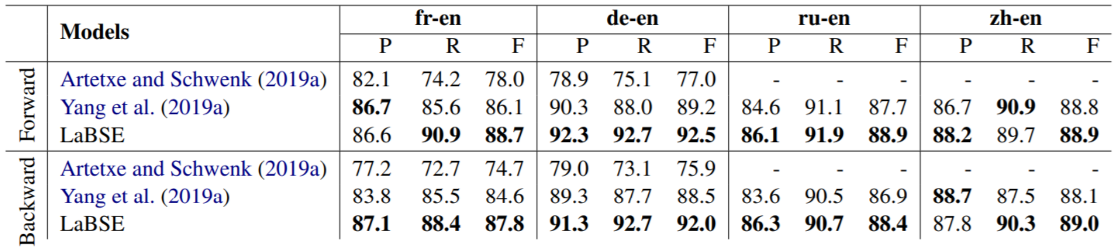
i \ j	0	1	2	3	4	5	6	7
0	1	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0
3	0	0	0	1	0	0	0	0
4	0	0	0	0	1	0	0	0
5	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	1

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 3: Negative sampling example in a dual encoder framework. The dot-product scoring function makes it efficient to compute the pairwise scores in the same batch with matrix multiplication. The value in the grids indicates the ground truth labels, with all positive labels located in diagonal grids. **[Left]**: The in-batch negative sampling in a single core; **[Right]**: *Synchronized multi-accelerator negative sampling* using n TPU cores and batch size 8 per core with examples from other cores are all treated as negatives.

Results



	Models	**fr-en**			**de-en**			**ru-en**			**zh-en**		
		P	R	F	P	R	F	P	R	F	P	R	F
Forward	Artetxe and Schwenk (2019a)	82.1	74.2	78.0	78.9	75.1	77.0	-	-	-	-	-	-
	Yang et al. (2019a)	**86.7**	85.6	86.1	90.3	88.0	89.2	84.6	91.1	87.7	86.7	**90.9**	88.8
	LaBSE	86.6	**90.9**	**88.7**	**92.3**	**92.7**	**92.5**	**86.1**	**91.9**	**88.9**	**88.2**	89.7	**88.9**
Backward	Artetxe and Schwenk (2019a)	77.2	72.7	74.7	79.0	73.1	75.9	-	-	-	-	-	-
	Yang et al. (2019a)	83.8	85.5	84.6	89.3	87.7	88.5	83.6	90.5	86.9	**88.7**	87.5	88.1
	LaBSE	**87.1**	**88.4**	**87.8**	**91.3**	**92.7**	**92.0**	**86.3**	**90.7**	**88.4**	87.8	**90.3**	**89.0**

Table 1: [P]recision, [R]ecall and [F]-score of BUCC training set score with cosine similarity scores. The thresholds are chosen for the best F scores on the training set. Following the naming of BUCC task (Zweigenbaum et al., 2018), we treat en as the target and the other language as source in forward search. Backward is vice versa.

Reason to read → The reason why I read this is to make the approaches language agnostic at least for the set of Indian Languages .

Directions to use → link