Crosslingual Embeddings are Essential in UNMT for Distant Languages: An English to IndoAryan Case Study

Tamali Banerjee ⋆ tamali@cse.iitb.ac.in
Department of Computer Science and Engineering, IIT Bombay, India.

Rudra Murthy V ⋆ rmurthyv@in.ibm.com
IBM Research Lab, India.

Pushpak Bhattacharyya pb@cse.iitb.ac.in
Department of Computer Science and Engineering, IIT Bombay, India.

Abstract

Recent advances in Unsupervised Neural Machine Translation (UNMT) have minimized the gap between supervised and unsupervised machine translation performance for closely related language-pairs. However, the situation is very different for distant language pairs. Lack of lexical overlap and low syntactic similarities such as between English and Indo-Aryan languages lead to poor translation quality in existing UNMT systems. In this paper, we show that initialising the embedding layer of UNMT models with cross-lingual embeddings shows significant improvements in BLEU score over existing approaches with embeddings randomly initialized. Further, static embeddings (freezing the embedding layer weights) lead to better gains compared to updating the embedding layer weights during training (non-static). We experimented using Masked Sequence to Sequence (MASS) and Denoising Autoencoder (DAE) UNMT approaches for three distant language pairs. The proposed cross-lingual embedding initialization yields BLEU score improvement of as much as ten times over the baseline for English-Hindi, English-Bengali, and English-Gujarati. Our analysis shows the importance of cross-lingual embedding, comparisons between approaches, and the scope of improvements in these systems.

1 Introduction

Unsupervised approaches to training a neural machine translation (NMT) system typically involve two stages: (i) Language Model (LM) pre-training and (ii) finetuning of NMT model using Back-Translated (BT) sentences. Training a shared encoder-decoder model on combined monolingual data of multiple languages helps the model learn better cross-lingual representations (Conneau et al., 2020; Wang et al., 2019). Fine-tuning the pre-trained model iteratively using Back-translated sentences helps further align the two languages closer in latent space and also trains an NMT system in an unsupervised manner.

Unsupervised MT has been successful for closely related languages (Conneau and Lample, 2019; Song et al., 2019). On the other hand, very poor translation performance has been reported

[*] The two authors contributed equally to this paper.

for distant language pairs (Kim et al., 2020a; Marchisio et al., 2020). Lack of vocabulary overlap and syntactic differences between the source and the target languages make the model fail to align the two language representations together. Recently, few approaches (Kulshreshtha et al., 2020; Wu and Dredze, 2020) take advantage of resources in the form of bilingual dictionary, parallel corpora, *etc.* to better align the language representations together during LM pre-training.

In this paper, we explore the effect of initialising the embedding layer with cross-lingual embeddings for training UNMT systems for distant languages. We also explore the effect of static cross-lingual embeddings (embedding are not updated during training) *v/s* non-static cross-lingual embeddings (embedding are updated during training). We experiment with two existing UNMT approaches namely, MAsked Sequence-to-Sequence (MASS) (Song et al., 2019) and a variation of Denoising Auto-Encoder (DAE) based UNMT approach (Artetxe et al., 2018c; Lample et al., 2018) for English to IndoAryan language pairs *i.e.* English-Hindi, English-Bengali, English-Gujarati.

The contribution of the paper is as follows:

1. We show that approaches initialized with cross-lingual embeddings significantly outperform approaches with randomly initialized embeddings.

2. We observe that the use of *static cross-lingual embeddings* leads to better gains compared to the use of *non-static* cross-lingual embeddings for these language-pairs.

3. We did a case study of UNMT for English-IndoAryan language pairs. For these language-pairs SOTA UNMT approaches perform very poorly.

4. We observed that DAE-based UNMT with crosslingual embeddings performs better than MASS-based UNMT with crosslingual embeddings for these language-pairs.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in detail. Then, we present our approach in Section 3. In Section 4, we outline the experimental setup and present the results of our experiments in Section 5. Finally, we conclude the paper and discuss future work in Section 6.

2 Related Work

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) typically needs a lot of parallel data to be trained on. However, parallel data is expensive and rare for many language pairs. To solve this problem, unsupervised approaches to train machine translation (Artetxe et al., 2018d; Lample et al., 2018; Yang et al., 2018) was proposed in the literature which uses only monolingual data to train a translation system.

Artetxe et al. (2018c) and Lample et al. (2018) introduced Denoising Auto-Encoder-iterative (DAE-iterative) UNMT which utilizes cross-lingual embeddings and trains a RNN-based encoder-decoder model (Bahdanau et al., 2015). Architecture proposed by Artetxe et al. (2018d) contains a shared encoder and two language-specific decoders while architecture proposed by Lample et al. (2018) contains a shared encoder and a shared decoder. In the approach by Lample et al. (2018), the training starts with word-by-word translation followed by denoising and backtranslation (BT). Here, noise in the input sentences in the form of shuffling of words and deletion of random words from sentences was performed.

Conneau and Lample (2019) (XLM) proposed a two-stage approach for training a UNMT system. The pre-training phase involves training of the model on the combined monolingual corpora of the two languages using Masked Language Modelling (MLM) objective (Devlin et al., 2019). The pre-trained model is later fine-tuned using denoising auto-encoding objective and backtranslated sentences. Song et al. (2019) proposed a sequence to sequence pre-training strategy. Unlike XLM, the pre-training is performed via MAsked Sequence to Sequence (MASS) objective. Here, random n-grams in the input are masked and the decoder is trained to generate the missing n-grams in the pre-training phase. The pre-trained model is later fine-tuned using backtranslated sentences.

Recently, Kim et al. (2020b) demonstrated that the performance of current SOTA UNMT systems is severely affected by language divergence and domain difference. The authors demonstrated that increasing the corpus size does not lead to improved translation performance. The authors hypothesized that existing UNMT approaches fail for distant languages due to lack of mechanism to bootstrap out of a poor initialization.

Recently, Chronopoulou et al. (2021) trained UNMT systems with 2 language pairs English-Macedonian (En-Mk) and English-Albanian (En-Sq) in low resource settings. These pairs achieved BLEU scores ranging from 23 to 33 using UNMT baseline XLM (Conneau and Lample, 2019) and RE-LM (Chronopoulou et al., 2020) systems. They showed further improvement up to 4.5 BLEU score when initialised embedding layer with crosslingual embedding. However, they did not explore the effect of initialising embedding layers on MASS, DAE-pretrained, and DAE-iterative approaches. Moreover, they did not experiment with language-pairs for which UNMT approaches with randomly initialised embedding layers fail completely even after training with a sufficient amount of monolingual data.

Additionally, there is some work on understanding multilingual language models and their effectiveness on zero-shot performance on downstream tasks (Pires et al., 2019; Kulshreshtha et al., 2020; Liu et al., 2020; Wang et al., 2020; Wu and Dredze, 2020). Here, the pre-trained multilingual language model is fine-tuned for the downstream NLP task in one language and tested on an unseen language (unseen during fine-tuning stage). While multilingual models have shown promising results on zero-shot transfer, the gains are limited for distant languages unless additional resources in the form of dictionary and corpora are used (Kulshreshtha et al., 2020; Wu and Dredze, 2020). Also, training a single model on unrelated languages might lead to negative interference (Wang et al., 2020).

3 Approaches

In this section, we explain different approaches used in our experiments. We use MASS (Song et al., 2019) and DAE based iterative approach similar to Lample et al. (2018) as our baseline models.

3.1 MASS UNMT

In MASS (Song et al., 2019), random n-grams in the input are masked and the model is trained to generate the missing n-grams in the pre-training phase. The pre-trained model is later fine-tuned using back-translated sentences. For every token, the input to the model is the summation of randomly initialised word embedding, positional encoding, and language code.

3.2 DAE UNMT

DAE UNMT approach is similar to the MASS UNMT approach with the difference being the pre-training objective. Here, we add random noise to the input sentence before giving it as input and the model is trained to generate the entire original sentence. Here, noise in the input sentences in the form of shuffling of words and deletion of random words from sentences was performed.

3.3 Cross-lingual Embedding Initialization

In both MASS and DAE UNMT approaches, the embedding layer is randomly initialized before the pre-training phase. We use Vecmap (Artetxe et al., 2018a) approach as a black-box to

Language	# train sentences
English (en)	54.3 M
Hindi (hi)	63.1 M
Bengali (bn)	39.9 M
Gujarati (gu)	41.1 M

Table 1: Monolingual Corpus Statistics in Million

Language-pair	# valid sentences	# test sentences
En - Hi	2000	3169
En - Bn	2000	3522
En - Gu	2000	4463

Table 2: Validation and Test Data Statistics

obtain cross-lingual embeddings. We then initialize the word-embedding layer with the cross-lingual embeddings obtained. During pre-training and fine-tuning, we have the opportunity to either *freeze* the embedding layer (static embeddings) or update them during training (non-static embeddings). We experiment with these two variations on both MASS and DAE approaches. We refer to MASS UNMT approach using static cross-lingual embeddings as *MASS + Static* and *MASS + Non-Static* for non-static cross-lingual embeddings. Similarly, We refer to DAE UNMT approach using static cross-lingual embeddings as *DAE + Static* and *DAE + Non-Static* for non-static cross-lingual embeddings.

3.4 DAE-iterative UNMT

Artetxe et al. (2018c) and Lample et al. (2018) proposed an approach based on Denoising Auto-Encoder and Back-Translation. Their approach trained the UNMT in one stage. During training, they alternated between denoising and back translation objectives iteratively. They initialised the embedding layer with cross-lingual embeddings and trained an RNN-based encoder-decoder model (Bahdanau et al., 2015). Architecture proposed by Artetxe et al. (2018d) contains a shared encoder and two language-specific decoders while architecture proposed by Lample et al. (2018) contains a shared encoder and a shared decoder, where all the modules are bi-LSTMs. We use Transformer-based architecture instead of bi-LSTM. In input, we do not add language code here. Similar to MASS and DAE, we experiment with using static and non-static cross-lingual embeddings.

4 Experimental Setup

We trained the models using 8 approaches for all language-pair out of which 3 approaches use DAE as LM pretraining, 3 approaches use MASS as LM pretraining, and the other two train DAE and BT simultaneously.

4.1 Dataset and Languages used

We use monolingual data of 4 languages *i.e.* English (en), Hindi (hi), Bengali (bn), Gujarati (gu). While English is of European language family, the other three languages are of Indo-Aryan language family. These three Indian languages follow Subject-Object-Verb word order. However, for English the word order is Subject-Verb-Object. We organise this experiment for distant language pairs with word-order divergence. Therefore, we pair English language with one of these three Indic languages resulting in three language-pairs, *i.e.* en-hi, en-bn, en-gu.

We use monolingual data provided by AI4Bharat (Kunchukuttan et al., 2020) dataset as training data. We use English-Indic validation and test data provided in WAT 2020 Shared task (Nakazawa et al., 2020) [*]. Details of our dataset used in this experiment are in Table 2.

[*] http://www.statmt.org/wmt20/translation-task.html

Language-pair	**en → x**		**x → en**	
	NN	CSLS	NN	CSLS
En - Hi	52.16 %	55.46 %	43.51 %	46.82 %
En - Bn	36.76 %	41.39 %	33.77 %	39.17 %
En - Gu	43.35 %	46.47 %	46.07 %	50.38 %

Table 3: Word-to-word translation accuracy using our crosslingual embeddings

4.2 Preprocessing

We have preprocessed the English corpus for normalization, tokenization, and lowercasing using the scripts available in *Moses* (Koehn et al., 2007) and the Indo-Aryan corpora for tokenization using *Indic NLP Library* (Kunchukuttan, 2020). For BPE segmentation we use *FastBPE*[†] jointly on the source and target data with number of merge operations set to 100k.

4.3 Word Embeddings

We use the BPE-segmented monolingual corpora to independently train the embeddings for each language using skip-gram model of *Fasttext*[‡] (Bojanowski et al., 2017). To map embeddings of the two languages to a shared space, we use *Vecmap*[§] to obtain cross-lingual embedding proposed by Artetxe et al. (2018b). We report the quality of the cross-lingual embeddings in Table 3 w.r.t. word-translation quality on MUSE data (Conneau et al., 2018) by nearest-neighbour and Cross-Domain Similarity Local Scaling (CSLS) approaches.

4.4 Network Parameters

We use MASS code-base [¶] and to tun our experiments. We train all the models with a 6 layer 8-headed transformer encoder-decoder architecture of dimension 1024. The model is trained using an epoch size of $0.2M$ steps and a batch size of 64 sentences (token per batch $3K$)). We use Adam optimizer with $beta_1$ set to 0.9, and $beta_2$ to 0.98, with learning rate to 0.0001. We pre-training for a total of 100 epochs and fine-tune for a maximum of 50. However, we stop the training if the model converges before the max-epoch is reached. The input to the model is a summation of word embedding and positional encoding of dimension 1024. In all our models, we drop the language code at the encoder side. For MASS pre-training we use word-mass of 0.5. Other parameters are default parameters given in the code-base. We do not search for optimised parameters, instead, we are looking for approaches that give decent results on most hyperparameters as hyperparameter tuning is very expensive.

4.5 Evaluation and Analysis

We report both BLEU scores as translation accuracy metric for these approaches. We additionally plot perplexity, accuracy, and BLEU scores for intermediate results of each model.

5 Result and Analysis

In this section, we present the results from our experiments and present a detailed analysis of the same.

[†] https://github.com/glample/fastBPE

[‡] https://github.com/facebookresearch/fastText

[§] https://github.com/artetxem/vecmap

[¶] https://github.com/microsoft/MASS

5.1 Results

The translation performance from our experiments is as shown in Table 4. We compared BLEU scores between models where embedding layers were initialised with cross-lingual embeddings and models where embedding layers were randomly initialised.

Initialising embedding layer with static cross-lingual embedding helps both MASS-based and DAE-based UNMT systems to learn better translations as seen from the table. Our results suggest that, freezing cross-lingual embeddings (static) during UNMT training results in better translation quality compared to the approach where cross-lingual embeddings are updated (non-static).

BLEU scores suggest that DAE objective based models surpass MASS objective based models for these language pairs. Though DAE-iterative models produce lower BLEU scores than *DAE Static* or *DAE Non-Static* models, the former approach gives better BLEU scores in less number of iterations as shown in Fig. 3.

For completeness, we compare the BLEU scores of the best UNMT model, *i.e. DAE Static*, with the best reported BLEU scores in WAT 2020 Shared Task (Nakazawa et al., 2020) reported by Yu et al. (2020) on the same test data in the supervised setting. The supervised approach uses parallel data in a multilingual setting. Their models reached high accuracy by improving baseline multilingual NMT models with Fast-align, Domain transfer, ensemble, and Adapter fine-tuning methods.

While our en-hi and en-gu models produce decent values of BLEU score, en-bn models produce low BLEU score. Intuitively, we assume language characteristics to be the reason behind it.

UNMT approaches	en → hi	hi → en	en → bn	bn → en	en → gu	gu → en
MASS	1.15	1.61	0.11	0.27	0.62	0.79
DAE	0.63	0.95	0.06	0.31	0.39	0.61
DAE-iterative Non-Static	5.37	6.63	1.66	4.19	3.12	5.98
MASS Non-Static	5.49	6.06	1.86	3.5	3.47	4.82
DAE Non-Static	7.65	8.85	2.35	4.67	4.55	6.84
DAE-iterative Static	7.96	9.09	2.88	5.54	5.63	8.64
MASS Static	5.5	6.49	2.09	4.7	4.13	6.09
DAE Static	**10.3**	**11.57**	**3.3**	**6.91**	**7.39**	**10.88**

Table 4: UNMT translation performance on distant languages, i.e. en-hi, en-bn, en-gu test sets (BLEU scores reported). The values marked in bold indicate the best score for a language pair.

System	en → hi	hi → en	en → bn	bn → en	en → gu	gu → en
Our best UNMT	10.3	11.57	3.3	6.91	7.39	10.88
SOTA Supervised NMT	24.48	28.51	19.24	23.38	14.16	30.26

Table 5: Comparison of results between our best unsupervised NMT models and SOTA supervised NMT models on WAT20 test data. Supervised NMT results are reported from Yu et al. (2020)

.

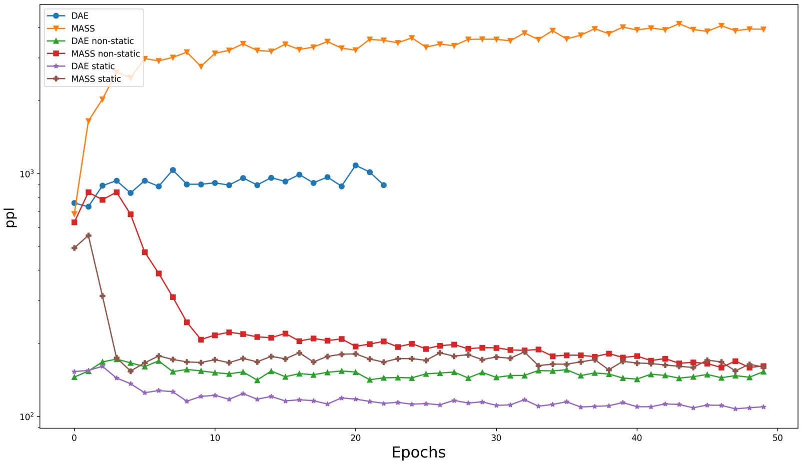

Figure 1: Change in Validation Set Translation Perplexity during Fine-tuning for English to Hindi Language pair

Hindi Source Word translation	आत्मनिर्भर बन रही है self-reliant becoming
English reference	it is becoming self reliant .
DAE	the same show is
MASS	employment back to the world
DAE Non-Static	it has become self - reliant
MASS Non-Static	resilient to the world
DAE Static	it is becoming self - sufficient
MASS Static	empowering the people

Figure 2: Example of a Hindi to English translation using various approaches

5.2 Analysis

We analyse the performance of our models by plotting translation perplexities on the validation set. Moreover, we manually analyse translation outputs and discuss them in this section.

5.2.1 Quantitative Analysis

In Fig. 1, we observe that for both MASS (baseline MASS) and DAE (baseline DAE) the plot of translation perplexity over epoch of finetuning stage increases rather than decreasing. On the other hand, when cross-lingual word embeddings are used the validation set translation perplexity decreases.

Among these embedding initialised models, we observe better convergence for models where embedding layers are frozen (static) than the models where embedding layers are updated (non-static). We also observe that the DAE-UNMT models converge better than MASS-UNMT models when initialized with cross-lingual embeddings.

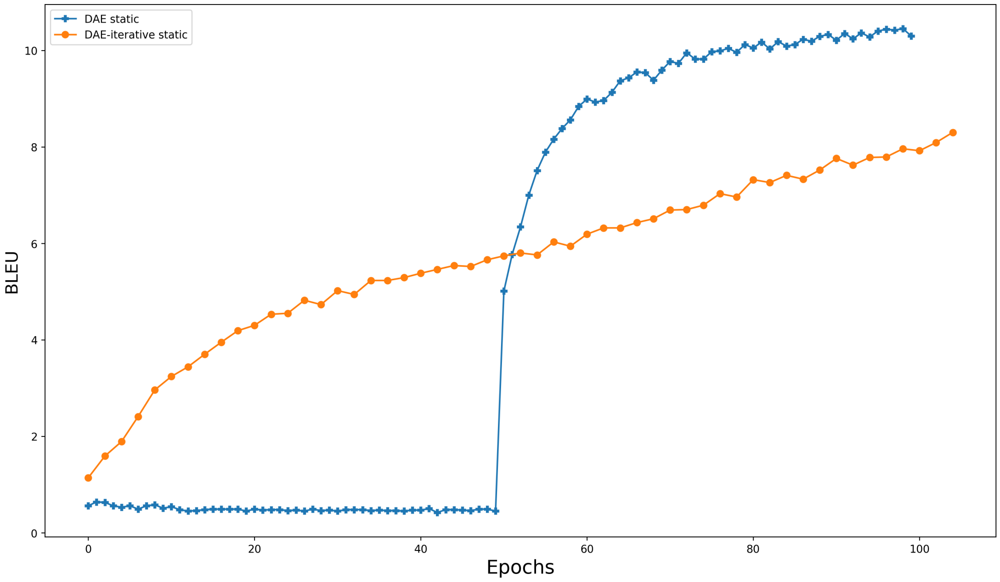

Figure 3: Comparison of Test Set BLEU Score for every epoch between DAE Static (DAE-pretrained UNMT) (both Pre-training and Fine-tuning) and DAE-iterative Static approach. Embedding layers of both the approaches are initialised with cross-lingual embedding and frozen during training. Language-pair: English-Hindi.

5.2.2 Qualitative Analysis

An example of a Hindi → English translation produced by various approaches is presented in Fig. 2. We observe the translation to be capturing the meaning of the source sentence when cross-lingual embeddings are used. However, we report some observations we found while analysing the translation outputs.

Lose of Phrasal Meaning We observe some translations where word meanings are prioritised over phrasal meaning. Fig. 4 shows such an example where dis-fluent translation is generated because of ignoring the phrasal meaning. Here, the model is unable to get the conceptual meaning of the sentence, instead translates words of the sentence literally.

Word Sense Ambiguity In Fig. 5 model fails to disambiguate word sense resulting in wrong translation. English word *'fine'* have different sense, *i.e.* beautiful and penalty. In this example, the model selects wrong sense of the word.

Scrambled Translation For many instances like Fig. 6, though the reference sentence and its corresponding generated sentences are formed with almost the same set of words, the sequence of words is different making the sentence lose its meaning. The error looks similar to the error addressed in Banerjee et al. (2019).

6 Conclusion

We show that existing UNMT methods such as DAE-based and MASS-based UNMT models fail for distant languages such as English to IndoAryan language pairs (*i.e.* en-hi, en-bn, en-gu). However, initialising the embedding layer with cross-lingual embeddings before Language Model (LM) pre-training helps the model train better UNMT systems for distant language pairs.

English Source	their hearts and my heart beat to the same rhythm .
Bengali reference	তাদের মনই আমার মন ।
English transliteration	tAdera manai AmAra mana |
Word translation	their mind my mind
System translation	তাঁদের হৃদয় এবং আমার হৃদয়ও একই ছন্দ মারিল ।
English transliteration	tA.Ndera hRRidaya ebaM AmAra hRRidayao ekai Chanda mArila |
Word translation	their heart and my heart same rhythm beat
English meaning	their hearts and my heart too beat to the same rhythm .

Figure 4: Example of a English to Bengali translation using DAE Static model

English Source	what a fine , purposeful message
Bengali reference	কত সুন্দর বার্তা ।
English transliteration	kata sundara bArtA |
Word translation	what a beautiful message .
System translation	কী একটা জরিমানা , purposeful বার্তা
English transliteration	kI ekaTA jarimAnA , purposeful bArtA
Word translation	what a penalty , purposeful message
English meaning	what a penalty/fine , purposeful message

Figure 5: Example of a English to Bengali translation using DAE Static model

English Source	they live in a parking shed with their family .
Bengali reference	তাঁরা সপরিবারে গাড়ি রাখার শেডের মধ্যে থাকেন ।
English transliteration	tA.NrA saparibAre gADai rAkhAra sheDera madhye thAkena |
Word translation	they with family parking shed inside lives
System translation	পার্কিং শেডের সঙ্গে বসবাস করে তাদের পরিবার ।
English transliteration	pArkiM sheDera sa Nge basabAsa kare tAdera paribAra |
Word translation	parking shed with live their family
English meaning	Their family live with parking shed

Figure 6: Example of a English to Bengali translation using DAE Static model

We also observe that static cross-lingual embedding gives better translation quality compared to non-static cross-lingual embeddings. For these distant language pairs, DAE objective based UNMT approaches produce better translation quality and converges better than MASS-based UNMT.

7 acknowledgements

The authors acknowledge the *IBM Research Cognitive Computing Cluster* service for providing resources that have contributed to the research results reported within this paper.

References

Artetxe, M., Labaka, G., and Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 789–798.

Artetxe, M., Labaka, G., and Agirre, E. (2018c). Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018d). Unsupervised neural machine translation. *Proceedings of the International Conference on Learning Representations (ICLR).*

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Banerjee, T., Murthy, V. R., and Bhattacharyya, P. (2019). Ordering matters: Word ordering aware unsupervised NMT. *CoRR*, abs/1911.01212.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Chronopoulou, A., Stojanovski, D., and Fraser, A. (2020). Reusing a pretrained language model on languages with limited corpora for unsupervised nmt. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711.

Chronopoulou, A., Stojanovski, D., and Fraser, A. (2021). Improving the lexical ability of pretrained language models for unsupervised neural machine translation.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. *Proceedings of the International Conference on Learning Representations (ICLR).*

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (2020). Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kim, Y., Graça, M., and Ney, H. (2020a). When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal. European Association for Machine Translation.

Kim, Y., Graça, M., and Ney, H. (2020b). When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Kulshreshtha, S., Garcia, J. L. R., and Chang, C. Y. (2020). Cross-lingual alignment methods for multilingual bert: A comparative study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 933–942.

Kunchukuttan, A. (2020). The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Kunchukuttan, A., Kakwani, D., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

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

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., et al. (2020). Overview of the 7th workshop on asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Wang, Z., Lipton, Z. C., and Tsvetkov, Y. (2020). On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Wang, Z., Xie, J., Xu, R., Yang, Y., Neubig, G., and Carbonell, J. G. (2019). Cross-lingual alignment vs joint training: A comparative study and A simple unified framework. *CoRR*, abs/1910.04708.

Wu, S. and Dredze, M. (2020). Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55.

Yu, Z., Wu, Z., Chen, X., Wei, D., Shang, H., Guo, J., Li, Z., Wang, M., Li, L., Lei, L., et al. (2020). Hw-tsc's participation in the wat 2020 indic languages multilingual task. In *Proceedings of the 7th Workshop on Asian Translation*, pages 92–97.