Scrambled Translation Problem: A Problem of Denoising UNMT

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

In this paper, we identify an interesting kind of error in the output of Unsupervised Neural Machine Translation (UNMT) systems like *Undreamt*[1]. We refer to this error type as *Scrambled Translation problem*. We observe that UNMT models which use *word shuffle* noise (as in case of Undreamt) can generate correct words, but fail to stitch them together to form phrases. As a result, words of the translated sentence look *scrambled*, resulting in decreased BLEU. We hypothesise that the reason behind *scrambled translation problem* is 'shuffling noise' which is introduced in every input sentence as a denoising strategy. To test our hypothesis, we experiment by retraining UNMT models with a simple *retraining* strategy. We stop the training of the Denoising UNMT model after a pre-decided number of iterations and resume the training for the remaining iterations- which number is also pre-decided- using original sentence as input without adding any noise. Our proposed solution achieves significant performance improvement UNMT models that train conventionally. We demonstrate these performance gains on four language pairs, *viz.*, English-French, English-German, English-Spanish, Hindi-Punjabi. Our qualitative and quantitative analysis shows that the retraining strategy helps achieve better alignment as observed by attention heatmap and better phrasal translation, leading to statistically significant improvement in BLEU scores.

1 Introduction

Training a machine translation system using only the monolingual corpora of the two languages was successfully demonstrated by (Artetxe et al., 2018c; Lample et al., 2018). They train the machine translation system using denoising auto-encoder (DAE) and backtranslation (BT) iteratively. Recently, pre-training of large language models (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020) using monolingual corpus is used to initialize the weights of the encoder-decoder models. These encoder-decoder models are later fine-tuned using backtranslated sentences for the task of Unsupervised Neural Machine Translation (UNMT). While we appreciate language model (LM) pre-training to better initialise the models, it is important to understand the shortcomings of earlier approaches. In this paper, we explore in this direction.

[1] https://github.com/artetxem/undreamt

We observe that the translation quality of undreamt models (Artetxe et al., 2018c) suffers partially due to wrong positioning of the target words in the translated sentence. For many instances, though the reference sentence and its corresponding generated sentence are formed with almost the same set of words, the sequence of words is different resulting in the sentence being ungrammatical and/or loss of meaning. This results in a difference in syntax and semantic rules. We define such generated sentences as **scrambled sentences** and the problem as **scramble translation problem**. Scrambled sentences can be either **disfluent** or **fluent-but-inadequate**. Here, if the LM decoder is not learnt well, we observe disfluent translations. If the LM decoder is learnt well, we observe fluent-but-inadequate translations. An example of fluent-but-inadequate translation will be *'leaving better kids for our planet'* instead of *'leaving better planet for our kids'*. Due to this phenomenon, during BLEU computation n-gram matching lessens, for $n > 1$. However, this error is absent in translation generated from recent state-of-the-art systems (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020).

We hypothesise, DAE introduces uncertainty to the previous UNMT (Lample et al., 2018; Artetxe et al., 2018c, 2019; Wu et al., 2019) models, specifically to the encoders. It has been observed that encoders are sensitive to the exact ordering of the input sequence (Michel and Neubig, 2018; Murthy V et al., 2019; Ahmad et al., 2019). By performing random word-shuffle in all the source sentences, encoder may lose important information about the sentence composition. The DAE fails to learn informative representation which affects the decoder resulting in wrong translations generated.

If our hypothesis is true, retraining these previous UNMT system models with noise-free sentences as input should resolve the problem for previous systems (Artetxe et al., 2018c; Lample et al., 2018). Moreover, using this retraining strategy will not benefit recent approaches (Conneau and Lample, 2019; Song et al., 2019) as they do not shuffle words of input sentence while training with back-translated data.

In this paper, we prove our hypothesis by showing that a simple **retraining strategy** mitigates the 'scrambled translation problem'. We observe consistent improvements in BLEU score and word-alignment over the denoising UNMT approach by Artetxe et al. (2018c) for four language pairs. We do not wish to beat the state-of-the-art UNMT systems with pre-training, instead, we demonstrate a limitation of previous denoising UNMT (Artetxe et al., 2018c; Lample et al., 2018) systems and prove why it happens.

2 Related Work

Neural machine translation (NMT) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) typically needs 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., 2018c; 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. (2018b) and Lample et al. (2018) introduced denoising-based U-NMT which utilizes cross-lingual embeddings and trains a RNN-based encoder-decoder model (Bahdanau et al., 2015). Architecture proposed by Artetxe et al. (2018c) 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. 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 ngrams in the input is masked and the decoder is trained to generate the missing ngrams in the pre-training phase. The pre-trained model is later fine-tuned using backtranslated sentences.

Murthy et al. (2019) demonstrated that LSTM encoders of the NMT system are sensitive to the word-ordering of the source language. They considered the scenario of zero-shot translation from language l_3 to l_2. They train a NMT system for $l_1 \rightarrow l_2$ languages and use l_1 - l_3 languages bilingual embeddings. This enables the trained model to perform zero-shot translation from $l_3 \rightarrow l_2$. However, if the word-order of the languages l_1 and l_3 are different, the translation quality from l_1 - l_3 is hampered.

Michel and Neubig (2018) have also made a similar observation albeit in the monolingual setting. They observe that accuracy of the machine translation system gets adversely affected due to noise in the input sentences. They discuss various sources of noise with one of them being word emission/insertion/repetition or grammatical errors. The lack of robustness to such errors could be attributed to the sequential processing of LSTM or Transformer encoders. As the encoder processes the input as a sequence and generates encoder representation at each time-step, such errors would lead to bad encoder representations resulting in bad translations generated. Similar observations have also been made by Ahmad et al. (2019) for cross-lingual transfer of dependency parsing. They observe that self-attention encoder with relative position representations is more robust to word-order divergence and enable better cross-lingual transfer for dependency parsing task compared to RNN encoders.

3 Baseline Approach

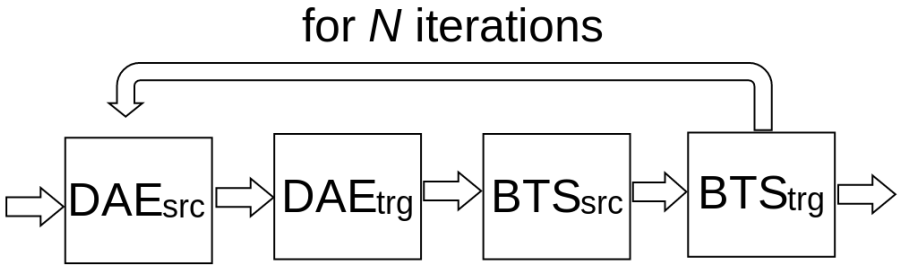


Figure 1: Our baseline training procedure: Undreamt. DAE_{src}: Denoising of source sentences; DAE_{trg}: Denoising of target sentences; BTS_{src}: Training with shuffled back-translated source sentences; BTS_{trg}: Training with shuffled back-translated target sentences.

We use Undreamt (Artetxe et al., 2018c) which is one of the previous UNMT approaches as the baseline for experimentation. Artetxe et al. (2018c) introduced denoising-based U-NMT which utilize cross-lingual embeddings and train a RNN-based encoder-decoder architecture Bahdanau et al. (2015). This architecture contains a shared encoder and two language-specific decoders. Training is a combination of denoising and back translation iteratively as shown in Fig. 1. By adding noise Artetxe et al. (2018c) meant shuffling of words of a sentence. Here, shuffling is performed by swapping neighboring words *l/2* times, where *l* is the number of words in the sentence. 4 sub-tasks of the training mechanism are listed below. (i) DAE_{src}: Denoising of source sentences in which we train shared-encoder, source-decoder, and attention with noisy

source sentence as input and original source sentence as output. (ii) DAE_{trg}: Denoising of target sentences which trains shared-encoder, target-decoder and attention with noisy target sentence as input and original target sentence as output. (iii) BTS_{src}: Training shared-encoder, target-decoder, and attention with shuffled back-translated source sentences as input and actual target sentences as output. (iv) BTS_{trg}: Training shared-encoder, source-decoder, and attention with shuffled back-translated target sentences as input and actual source sentences as output. Here, shuffling is performed by swapping neighboring words *l/2* times, where *l* is the number of words in the sentence.

For completeness, we also experimented with XLM UNMT (Conneau and Lample, 2019) with initialise the model with MLM objective followed by finetuning it with DAE and BT iteratively. In this approach, they do not add noise with the input sentence while training with backtranslated data.

4 Proposed Retraining Strategy

Our proposed strategy to train a denoising-based UNMT system consists of two phases. In the first phase, we proceed with training using denoised sentences similar to the baseline system (Artetxe et al., 2018c) for *M* number of iterations. Adding random shuffling in the input side, however, could introduce uncertainty to the model leading to inconsistent encoder representations. To overcome this, in the second phase, we retrain the model with simple AE and on-the-fly BT using sentences with the correct ordering of words for *(N-M)* iterations as shown in Fig. 2. Here, N is the total number of iterations and $M < N$. More concretely, this training approach consists of 4 more sub-processes other than the 4 subprocesses of the baseline system. These are: (v) AE_{src}: Auto-encoding of source sentences in which we train shared-encoder, source-decoder, and attention. (vi) AE_{trg}: Auto-encoding of target sentences in which we train shared-encoder, target-decoder, and attention. (vii) BT_{src}: Training shared-encoder, target-decoder, and attention with back-translated source sentences as input and actual target sentences as output. (viii) BT_{trg}: Training shared-encoder, source-decoder, and attention with back-translated target sentences as input and actual source sentences as output. The second phase ensures that the encoder learns to generate context representation with information about the correct ordering of words. For XLM (Conneau and Lample, 2019), we add these 4 subprocesses only with fine-tuning step. We do not change anything in LM pretraining step.

5 Experimental Setup

We test our hypothesis with undreamt as a previous approach and XLM as a SOTA approach. We applied our *retraining strategy* on both the approaches and observed the result.

For undreamt, we have used monolingual data of six languages, *i.e.* English (en), French (fr), German (de), Spanish (es), Hindi (hi), and Punjabi (pa). Among these languages, Hindi and Punjabi are of SOV word-order where the other four languages are of SVO word order. In our experiments, we choose language-pairs such that the word-order of source language matches with that of target language. We have used the NewsCrawl corpora for en, fr, de of WMT14, and for es of WMT13. For hi-pa, we use Wikipedia dumps of the august 2019 snapshot for training. The en-fr and en-de models are tested using WMT14 test-data and en-es models using WMT13 test-data, and hi-pa models using ILCI test data (Jha, 2010).

We have preprocessed the corpus for normalization, tokenization and lowercasing using the scripts available in *Moses* (Koehn et al., 2007) and *Indic NLP Library* (Kunchukuttan, 2020), for BPE segmentation using *subword-NMT* (Sennrich et al., 2016) with number of merge operations set to 50k.

We use the monolingual corpora to independently train the embeddings for each language using skip-gram model of *word2vec* (Mikolov et al., 2013). To map embeddings of two languages

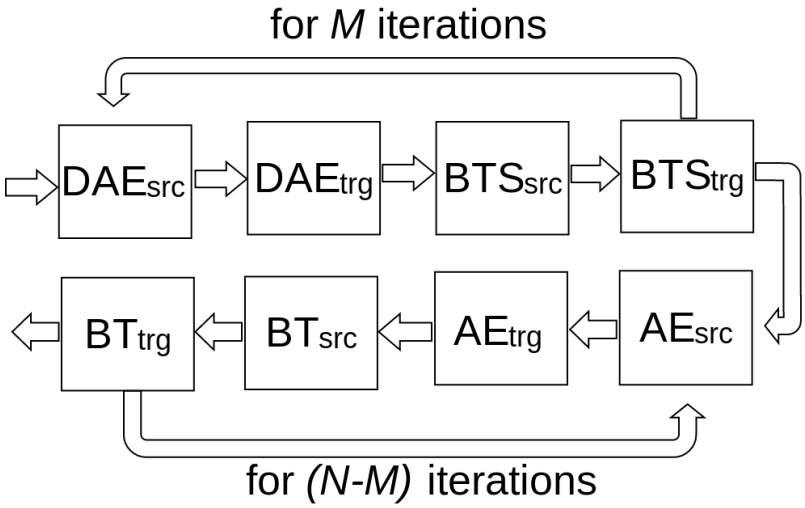


Figure 2: Workflow of Proposed training procedure. DAE_{src}: Denoising of source sentences; DAE_{trg}: Denoising of target sentences; BTS_{src}: Training with shuffled back-translated source sentences; BTS_{trg}: Training with shuffled back-translated target sentences; AE_{src}: Autoencoding of source sentences; AE_{trg}: Autoencoding of target sentences; BT_{src}: Training with shuffled back-translated source sentences; BT_{trg}: Training with shuffled back-translated target sentences.

to a shared space, we use *Vecmap*[2] by Artetxe et al. (2018a).

We use *undreamt*[3] tool to train the UNMT system proposed by Artetxe et al. (2018c). We train the baseline model untill convergence and noted the number of steps N required to reach convergence. We now train our proposed system for $N/2$ steps and re-train the model after removing denoising noise for the remaining $N/2$ steps. They converge between 500k to 600k steps depending on the language pairs. Further details of dataset and network parameters are available in Appendix.

We also report results on *XLM*[4] approach (Conneau and Lample, 2019). XLM employs two-stage training of UNMT model. The pre-training stage trains encoder and decoder with masked language modeling objective. The retraining stage employs denoising along with iterative back-translation. However, XLM uses a different denoising (word shuffle) mechanism compared to Artetxe et al. (2018c). We replace the denoising mechanism by Conneau and Lample (2019) with the denoising mechanism used by Artetxe et al. (2018c). We use the pre-trained models for English-French, English-German, and English-Romanian provided by Conneau and Lample (2019). We retrain the XLM model until convergence using the denoising approach which makes the baseline system. We later retrain the pre-trained XLM model using our proposed approach where we remove the denoising component after $N/2$ steps.

We report both BLEU scores and n-gram BLEU scores using *multi-bleu.perl* of Moses. We have tested statistical significance of BLEU improvements (Koehn, 2004). To analyse the systems, we have produced heatmaps of attention generated by the models.

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

[3] https://github.com/artetxem/undreamt

[4] https://github.com/facebookresearch/XLM

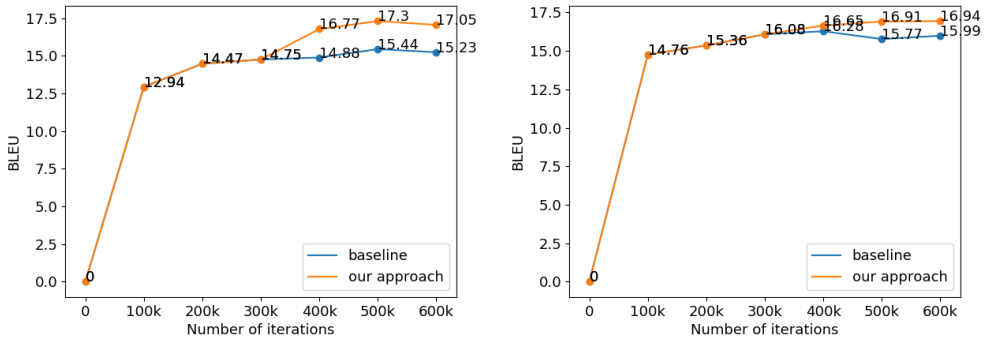
Language Pairs	Baseline (Undreamt)	Retrain with AE+BT†
en→fr	15.23	**17.05**
fr→en	15.99	**16.94**
en→de	6.69	**8.03**
de→en	10.67	**11.66**
en→es	15.09	**16.97**
es→en	15.33	**17.12**
hi→pa	22.39	**28.61**
pa→hi	28.38	**33.59**

(a) The translation performance using Undreamt-baseline and Undreamt-retraining on en-fr, en-de, en-es, hi-pa test sets (BLEU scores reported).

Language Pairs	Baseline (XLM)	Retrain with AE+BT
en→fr	**33.24**	31.94
fr→en	**31.34**†	30.79
en→de	25.06	25.02
de→en	30.53	30.34
en→ro	31.37	**31.72**
ro→en	29.01	**29.96**†

(b) The translation performance using XLM-baseline and XLM-retraining on en-fr, en-de, en-ro test sets (BLEU scores reported).

Table 1: The Translation performance using the Baseline approach and our Approach. Trained for a total of N iterations for all approaches. *Undreamt* and *XLM* results are results from our replication using the code provided by the authors. † indicates statistically significant improvements using paired bootstrap re-sampling (Koehn, 2004) for a p-value less than 0.05 .



(a) English → French

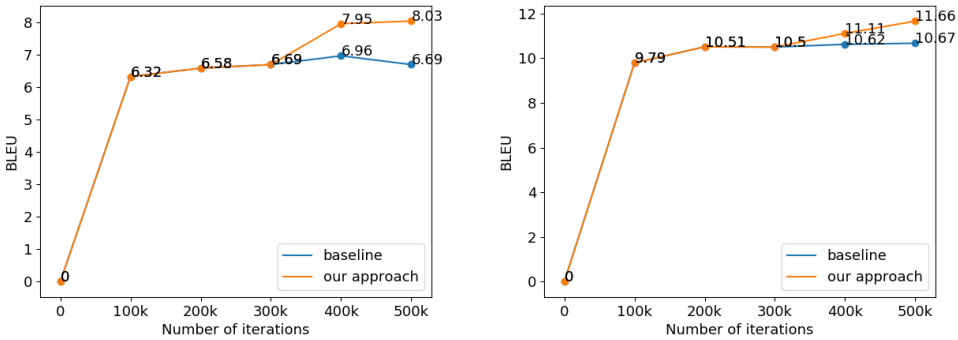
(b) French → English

Figure 3: Change in translation accuracy using undreamt-baseline vs. our approach with increasing number of iterations for English-French (BLEU scores reported).

6 Results and Analysis

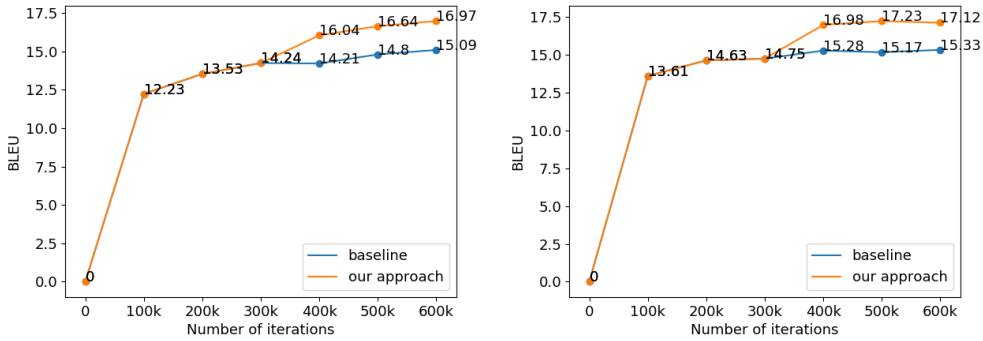
Table 1 reports BLEU score of the trained models using the undreamt (Artetxe et al., 2018c) and XLM (Conneau and Lample, 2019) and retraining them with our approach. *Undreamt* and *XLM* results are results from our replication using the code provided by the authors. In Table 1a we observe that the proposed re-training strategy of AE used in conjunction with BT results in statistically significant improvements (p-value < 0.05) across all language pairs when compared to the undreamt baseline approach (Artetxe et al., 2018c).

We report results on XLM (Conneau and Lample, 2019) with our *retraining* approach in Table 1b. XLM is one of the state-of-the-art (SOTA) UNMT approaches for these language pairs. The approach by XLM (Conneau and Lample, 2019) does not add noise to the input backtranslated sentence during training. Therefore, our retraining strategy does not benefit here.



(a) English → German (b) German → English

Figure 4: Change in translation accuracy using undreamt-baseline vs. our approach with increasing number of iterations for English-German (BLEU scores reported).

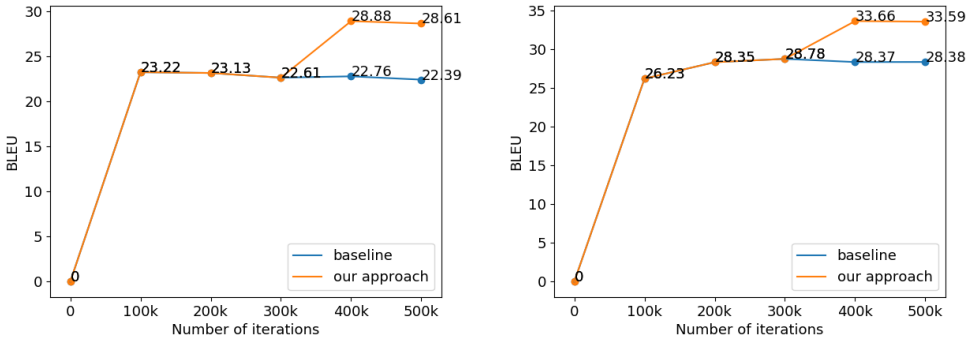


(a) English → Spanish (b) Spanish → English

Figure 5: Change in translation accuracy using undreamt-baseline vs. our approach with increasing number of iterations for English-Spanish (BLEU scores reported).

Language Pairs	Δ **BLEU**-1	Δ **BLEU**-2	Δ **BLEU**-3	Δ **BLEU**-4
en→fr	0.00	4.50	8.85	11.67
fr→en	2.17	5.53	7.48	10.90
en→de	17.44	11.71	17.07	25.00
de→en	1.75	6.87	12.12	13.33
en→es	1.75	6.88	12.04	20
es→en	3.20	9.13	14.85	21.15
hi→pa	7.49	24.48	32.71	46.39
pa→hi	4.30	15.89	24.12	30.56

Table 2: Improvements in n-BLEU (represented in %) on using our approach over baseline for en-fr, en-de, en-es, hi-pa test sets.



(a) Hindi → Punjabi (b) Punjabi → Hindi

Figure 6: Change in translation accuracy using undreamt-baseline vs. our approach with increasing number of iterations for Hindi-Punjabi (BLEU scores reported).

German	der us-senat genehmigte letztes jahr ein 90 millionen dollar teures pilotprojekt , das 10.000 autos umfasst hätte .
English reference	the u . s . senate approved a $ 90 - million pilot project last year that would have involved about 10,000 cars .
Artetxe et al. 2018	the u . s . district of the last $ 90 million a year , it would have 10,000 cars .
Our approach	the u . s . district last year approved 90 million initiative that would have included 10,000 cars .

Figure 7: Sample translation of German → English translation models.

Punjabi	ਸੁੱਕੇ	ਅੰਗੂਰ	ਜਾਂ ਫਿਰ	ਕਿਸ਼ਮਿਸ਼	ਵਿਚ	ਪਾਣੀ	ਦੀ	ਮਾਤਰਾ	੧੫	ਪ੍ਰਤੀਸ਼ਤ	ਹੁੰਦੀ ਹੈ	।
(Word transliteration)	suke	agUra	jAM phira	kisamisa	vicha	pANI	dI	mAtarA	15	pratIsata	hudI hai	|
(Word-to-word translation)	dry	grapes	or	raisins	in	water	of	quantity	15	percent	is	.
(Sentence translation)	Dried grapes or raisins have 15 percent water content .											
Hindi reference	सूखे	अंगूर	या फिर	किशमिश	में	पानी	की	मात्रा	१५	प्रतिशत	होती है	।
(Word transliteration)	sUkhe	aMgUra	yA phira	kishamisha	meM	pAnI	kI	mAtrA	15	pratishata	hotI hai	|
(Word-to-word translation)	dry	grapes	or	raisins	in	water	of	quantity	15	percent	is	.
Artetxe et al. 2018	अंगूर	या फिर	अंगूर	में	फिर से	पानी	की	मात्रा	१२	प्रतिशत	होती है	।
(Word transliteration)	aMgUra	yA phira	aMgUra	meM	phira se	pAnI	kI	mAtrA	12	pratishata	hotI hai	|
(Word-to-word translation)	grapes	or	grapes	in	again	water	of	quantity	12	percent	is	.
(Sentence translation)	The amount of water in the grape or grape again is 12 percent .											
Our approach	सूखे	अंगूर	या फिर	मालवण	में	पानी	की	मात्रा	१२	प्रतिशत	होती है	।
(Word transliteration)	sUkhe	aMgUra	yA phira	mAlavaNa	meM	pAnI	kI	mAtrA	12	pratishata	hotI hai	|
(Word-to-word translation)	dry	grapes	or	Malavan	in	water	of	quantity	12	percent	is	.
(Sentence translation)	Dried grapes or Malavan have 12 percent water content .											

Figure 8: Sample translation of Punjabi→Hindi translation models.

Spanish	el anuncio del probable descubrimiento del bosón de higgs generó una gran conmoción el verano pasado , y con razón .
English reference	the announcement of the probable discovery of the higgs boson created quite a stir last summer , and with good reason .
Artetxe et al. 2018	the likely announcement of the discovery of the higgs boson triggered a major shock last summer , and with reason .
Our approach	the announcement of the likely discovery of the higgs boson generated a major shock last summer , and with reason .

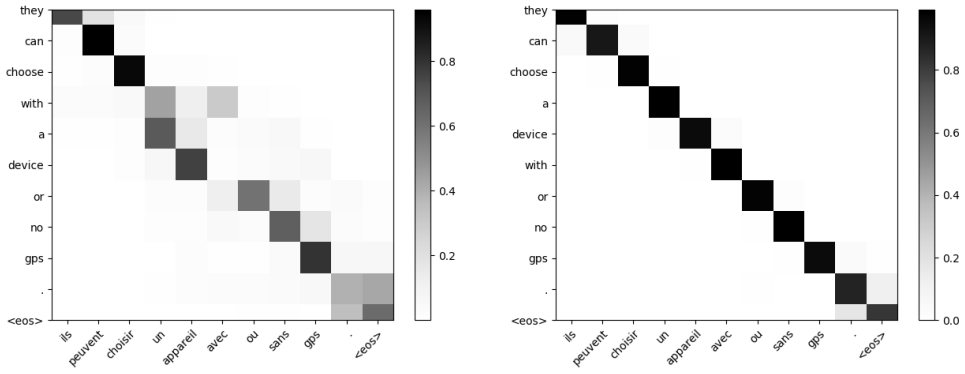
Figure 9: Sample translation of Spanish → English translation models.

We also observe robustness of the pre-trained language models to the scrambled translation problem.

Fig. 3, 4, 5 and 6 show changes in BLEU scores of intermediate UNMT models with

English	in india , china and many other countries , people work ten to twelve hours a day .
French reference	en inde , en chine et dans plein d' autres pays , on travaille dix à douze heures par jour .
Artetxe et al. 2018 (Google translation)	en inde , chine et autres pays , les autres gens travaillent à quinze heures à un jour . In India, China and other countries, other people work from fifteen to one.
Our approach (Google translation)	en inde , en chine et de nombreux autres pays , les gens travaillent quinze à douze heures un jour . In India, China and many other countries, people work fifteen to twelve hours a day .

Figure 10: Sample translation of English → French translation models.



(a) Using baseline approach

(b) Using our approach

Figure 11: Attention heatmaps of a French→English translation.

increasing number of iterations on test-data. We observe that our proposed approach leads to increase in BLEU score in the re-training phase as the denoising strategy is removed. The baseline system suffers from drop in BLEU score due to denoising strategy introducing ambiguity into the model.

6.1 Quantitative analysis

We hypothesize that the baseline UNMT model using DAE is able to generate correct word translation but fails to stitch them together to generate phrases. To validate the hypothesis, we calculate the percentage improvement on using our approach over the baseline system in terms of individual n-gram (n=1,2,3,4) specific BLEU scores for each language-pair and a particular value of *n*. The results presented in Table 2 indicate that our method achieves higher improvements in n-gram BLEU for higher $n-$grams $(n > 1)$ compared to the improvement in n-gram BLEU for lower values of *n*, indicating better phrasal matching. This could be attributed to the proposed approach not suffering from the *scrambled translation problem* introduced by the DAE.

6.2 Qualitative analysis

We observe several instances where our proposed approach results in better translations compared to the baseline. On manual analysis of translation outputs generated by the baseline system, we have found out some instances of *scrambled translation problem*.

Due to uncertainty introduced by shuffling of words before training, the baseline model

chooses to generate sentences that are more acceptable by a language model. Fig 7 shows such an example in our test data. Here, two German phrases *'ein 90 millionen'* (*'a 90 million'*) and *'letztes jahr'* (*'last year'*) are mixed up and translated as *'last $ 90 million a year'* in English. However, our approach handled the issue correctly.

Fig 8 shows an example of a situation where the baseline model prefers to generate a word in multiple probable positions. Here, the source Punjabi sentence consists of a phrase *'jAM phira'* (*'or'*) meaning *'yA phira'*(*'or'*) in Hindi. In the translation produced by the baseline model, the correct phrase is generated along with the word *'phira'* wrongly occurring again forming another phrase *'phira se'* (*'again'*). Note that, both the phrases are commonly used in Hindi. In Fig 9, the model trained on baseline system produced the word *'likely'*, which is a synonym of 'probably', in the wrong position. In Fig 10, the model trained on baseline system produced the word *'autres'*(*'other'*) in the multiple positions.

Attention Analysis: Attention distributions generated by our proposed systems have lesser confusion when compared with the attention distribution generated by baseline systems, as shown in Heatmaps of Fig. 11. Production of word-aligned attention distribution was easy for the attention models, which we retrained on sentences without noise.

7 Conclusion and Future work

In this paper, we addressed 'scrambled translation problem', a shortcoming of previous denoising-based UNMT approaches like *UndreaMT* approach (Artetxe et al., 2018c; Lample et al., 2018). We demonstrated that adding shuffling noise to all input sentences is the reason behind it. Our simple *retraining* strategy, *i.e.* retraining the trained models by removing the denoising component from auto-encoder objective (AE), results in significant improvements in BLEU scores for four language pairs. We observe larger improvements in n-gram specific BLEU scores for higher value of *n* indicating better phrasal translations. We also observe robustness of the pre-trained language models to the scrambled translation problem. We would also like to explore applicability of our approach in other ordering-sensitive DAE-based tasks.

References

Ahmad, W., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., and Peng, N. (2019). On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

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)*, volume 1, pages 789–798.

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

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018c). 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)*.

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.

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

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.

Jha, G. N. (2010). The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004*.

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.

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

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.

Michel, P. and Neubig, G. (2018). MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Murthy, R., Kunchukuttan, A., and Bhattacharyya, P. (2019). Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. 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 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

Murthy V, R., Kunchukuttan, A., Bhattacharyya, P., et al. (2019). Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

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

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.

Wu, J., Wang, X., and Wang, W. Y. (2019). Extract and edit: An alternative to back-translation for unsupervised neural machine translation. *Proceedings of the 57th Annual Meeting of the 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.