# Comparing DAE-based and MASS-based UNMT: Robustness to Word-Order Divergence in English→Indic Language Pairs

**Tamali Banerjee**
IIT Bombay
tamali@cse.iitb.ac.in

**Rudra Murthy V.**
IBM research lab
rmurthyv@in.ibm.com

**Pushpak Bhattacharyya**
IIT Bombay
pb@cse.iitb.ac.in

## Abstract

We test the robustness of state-of-the-art Unsupervised NMT (UNMT) approaches (*i.e.*, MASS-based UNMT and DAE-based UNMT) to word-order divergence between source and target languages. We investigate this by comparing two models for each of the two approaches, *i.e.*, (i) model trained on language pairs with different word-orders, and (ii) model trained on the same language pairs with source language re-ordered to match the word-order of the target language. Ideally, UNMT approaches that are robust to word-order divergence should exhibit no visible performance difference between the two configurations. Our study focuses on five English→Indic language pairs (*i.e.*, en-hi, en-bn, en-gu, en-kn, and en-ta) with SVO source word-order and SOV target word-order. Our findings show that DAE-based UNMT consistently outperforms MASS-based UNMT in translation accuracy for these language pairs. Bridging the word-order gap through reordering improves the accuracy of MASS-based UNMT models but does not improve DAE-based UNMT models. This suggests that DAE-based UNMT is more robust to word-order divergence.

## 1 Introduction

Unsupervised Neural Machine Translation (UNMT) shows promising results for closely related language-pairs (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020), but it faces significant challenges when dealing with language-pairs that have distinct word orders. In this paper, we test the robustness of word-order divergence in state-of-the-art UNMT systems. Specifically, we test MASS-based UNMT (Song et al., 2019; Banerjee et al., 2021) (which does not have shuffling noise) and DAE-based UNMT systems (Liu et al., 2020; Banerjee et al., 2021) (which has shuffling noise) on language pairs with different word-orders, *i.e.*, English (SVO) → Indic (SOV) language pairs.

To test the robustness, we compare these UNMT models trained on (i) original data and (ii) re-ordered data (where the source sentences are re-ordered to match the target language word-order). Word-order divergence is present in the former case, while word-order divergence is bridged in the latter case. A UNMT system that is robust to word-order divergence should not exhibit significant performance differences between these two cases.

Our contributions encompass two key findings: (i) DAE-based UNMT demonstrates greater robustness in managing word-order differences between languages compared to MASS-based UNMT, and (ii) in the majority of language pairs, the UNMT model trained with the DAE approach using original data produces translations of higher quality than other models.

## 2 Related work

Previous research has addressed lexical divergence between languages in NLP (Bhattacharyya, 2012) through various techniques (Chronopoulou et al., 2021; Banerjee et al., 2021; Khatri et al., 2021). However, the impact of word-order divergence on Unsupervised Neural Machine Translation (UNMT) remains unexplored. The study by Sun et al. (2021) examined the iterative UNMT approach proposed by Lample et al. (2018) and was found to be sensitive to word-order divergence.

Re-ordering addresses word-order differences in Machine Translation. While Du and Way (2017) found it unnecessary for NMT, Zhao et al. (2018) improved translation. In transfer learning, Murthy V et al. (2019) enhanced results by re-ordering source language sentences before training.

## 3 Approaches used

We use MASS-static and DAE-static methods (Banerjee et al., 2021) to address lexical divergence. These approaches initialize the models' embedding layers with unsupervised cross-lingual embeddings and keep the embedding layers static throughout UNMT training.

### 3.1 Language-model objectives

In the MASS (MAsked Sequence to Sequence) objective (Song et al., 2019), a random n-gram token of size $k$ (where $k$ is half of the sentence length) is selected in the input sentence. Within that fragment, 80% of the tokens are masked, 10% are replaced by random tokens, and the remaining 10% are left unchanged. The model is then trained to generate the missing n-grams. In the DAE (Denoising Auto-Encoder) objective, we introduce random noise to the input sentence, and the model is trained to reconstruct the original sentence. We employ word shuffle, word mask, and word deletion noise following the approach by Artetxe et al. (2018).

### 3.2 UNMT with re-ordering

For training a UNMT model with re-ordered data, we align the source sentence word-order with the target language. The model is trained using the re-ordered source and target monolingual data. During testing, source test sentences are also re-ordered before being inputted into the model. However, these models are not suitable for target→source translation, as they generate source sentences in the re-ordered form.

## 4 Experimental setup

Our experiment comprises four sets of UNMT models: two trained using MASS-static, and the other two trained using DAE-static. For each UNMT approach, we train one model on original data and another model on re-ordered data.

### 4.1 Language and datasets

We use six languages: English (en), Hindi (hi), Bengali (bn), Gujarati (gu), Kannada (kn), and Tamil (ta). Among these languages, English follows the Subject-Verb-Object (SVO) word-order, while the other five Indian languages have the Subject-Object-Verb (SOV) word-order. In our experiment, we focus on five language pairs: en→hi, en→bn, en→gu, en→kn, and en→ta. We use monolingual data provided by IndicCorp dataset (Kakwani et al., 2020) as training data. We use English-Indic validation and test data provided in WAT 2021 Shared task (Nakazawa et al., 2021).

### 4.2 Preprocessing tools

We have tokenised the English corpus using *Moses* (Koehn et al., 2007) and the Indic corpora using *Indic NLP Library* (Kunchukuttan, 2020). We use Generic rules of CFILT-pre-order (Chatterjee et al., 2014) for re-ordering English sentences to match word-order of Indic languages. We use *FastBPE*[1] jointly on the source and target data with the number of merge operations set to 100k. We follow the crosslingual embedding setup given by Banerjee et al. (2021).

### 4.3 Reordering noise removal

To ensure a fair comparison, we excluded sentences from both the original and re-ordered data that resulted in parse errors during the reordering process. We applied the same exclusion criteria to the valid and tested parallel data, removing their translations as well. As a result, the data size slightly decreased. The specifics of the remaining data are in Table 1.

### 4.4 Network and evaluation

We use MASS code-base (Song et al., 2019) and their default settings. The model is trained using an epoch size of $0.2M$ steps and a batch size of 64 sentences (token per batch $3K$)). For each of pretraining and finetuning steps, we train the models for 50 epochs maximum. However, we stop the training if the model converges before the max-epoch is reached based on validation split loss. For MASS pretraining, we use word-mass of $0.5$. For DAE pretraining, we use word-shuffle 3, word-dropout $0.1$, and word-blank $0.1$. We report BLEU (Papineni et al., 2002) and CHRF (Popović, 2016) (beneficial for morphologically rich languages) scores of the systems using SacreBLEU (Post, 2018).

## 5 Result and analysis

Table 2 presents the BLEU and CHRF scores of our UNMT models with and without re-ordering (R). Additionally, we indicate the impact of incorporating re-ordering on the translation quality, whether it led to improvement or degradation in terms of BLEU and CHRF scores. We exclude the results in the target→source direction for re-ordered UNMT models, as they generate trans-

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

| Language | # train sentences |
|---|---|
| English (en) | 52.2 M |
| Hindi (hi) | 63.1 M |
| Bengali (bn) | 39.9 M |
| Gujarati (gu) | 41.1 M |
| Kannada (kn) | 53.3 M |
| Tamil (ta) | 31.5 M |

| Language-pair | # valid sentences | # test sentences |
|---|---|---|
| en - hi | 711 | 1781 |
| en - bn | 711 | 1781 |
| en - gu | 711 | 1781 |
| en - kn | 711 | 1781 |
| en - ta | 711 | 1781 |

Table 1: Dataset statistics after noise removal

| Evaluation metrics | UNMT approaches | Translation accuracies on different language-pairs | | | | | |
|---|---|---|---|---|---|---|---|
| | | en - hi | | en - bn | | en - gu | |
| | | S → T | T → S | S → T | T → S | S → T | T → S |
| BLEU | MASS-static | 14.16 | 14.03 | 1.51 | 2.77 | 5.31 | 6.25 |
| | MASS-static + R | 14.63 (↑ 0.47) | - | 3.04 (↑ 1.53) | - | 8.62 (↑ 3.31) | - |
| | DAE-static | **21.03** | **21.89** | 2.88 | **4.39** | **10.60** | **14.78** |
| | DAE-static + R | 15.22 (↓5.81) | - | **3.27** (↑ 0.39) | - | 8.72 (↓ 1.88) | - |
| CHRF | MASS-static | 39.45 | 46.07 | 25.39 | 29.56 | 31.08 | 36.87 |
| | MASS-static + R | 41.77 (↑ 2.32) | - | 28.85 (↑ 3.46) | - | 37.45 (↑ 6.37) | - |
| | DAE-static | **45.63** | **52.21** | 28.36 | **34.61** | **38.55** | **45.74** |
| | DAE-static + R | 42.18 (↓ 3.45) | - | **29.65** (↑ 1.29) | - | 37.58 (↓ 0.97) | - |

| | | en - kn | | en - ta | |
|---|---|---|---|---|---|
| | | S → T | T → S | S → T | T → S |
| BLEU | MASS-static | 3.08 | 5.11 | 1.81 | 2.70 |
| | MASS-static + R | **4.48** (↑ 1.4) | - | 2.57 (↑ 0.76) | - |
| | DAE-static | 4.42 | **9.40** | 2.52 | **4.77** |
| | DAE-static + R | 4.18 (↓ 0.24) | - | **2.65** (↑ 0.13) | - |
| CHRF | MASS-static | 30.73 | 33.83 | 31.40 | 28.73 |
| | MASS-static + R | **36.47** (↑ 5.74) | - | 34.34 (↑ 2.94) | - |
| | DAE-static | 35.89 | **40.33** | 33.00 | **34.48** |
| | DAE-static + R | 36.46 (↑ 0.57) | - | **35.52** (↑ 2.52) | - |

Table 2: Translation accuracies of UNMT models with/without re-ordering (R) in both directions (S→T and T→S). Bold values indicate the best scores, values in parenthesis denote the improvement/degradation in BLEU/CHRF compared to the model above them.

| | |
|---|---|
| **English source sentence** | **We** need to change this mindset . |
| **Reordered English source sentence** | **We** this mindset change to need . |
| **Hindi reference** | हमें इस सोच को बदलने की ज़रूरत है ।<br>**hameM** isa socha ko badalane kI jarUrata hai \| |
| **Translation using DAE-static** | हमें इस मानसिकता को बदलना होगा ।<br>**hameM** isa mAnasikatA ko badalanA hogA \|<br>We need to change this mindset . |
| **Translation using reordered-DAE-static** | हम मानसिकता में बदलाव की जरूरत नहीं है ।<br>**hama** isa mAnasikatA meM badalAva kI jarUrata nahIM hai \|<br>We don't need a change in mindset [*case-marker is missing in output*] |

Figure 1: Translation example where re-ordering creates ambiguity

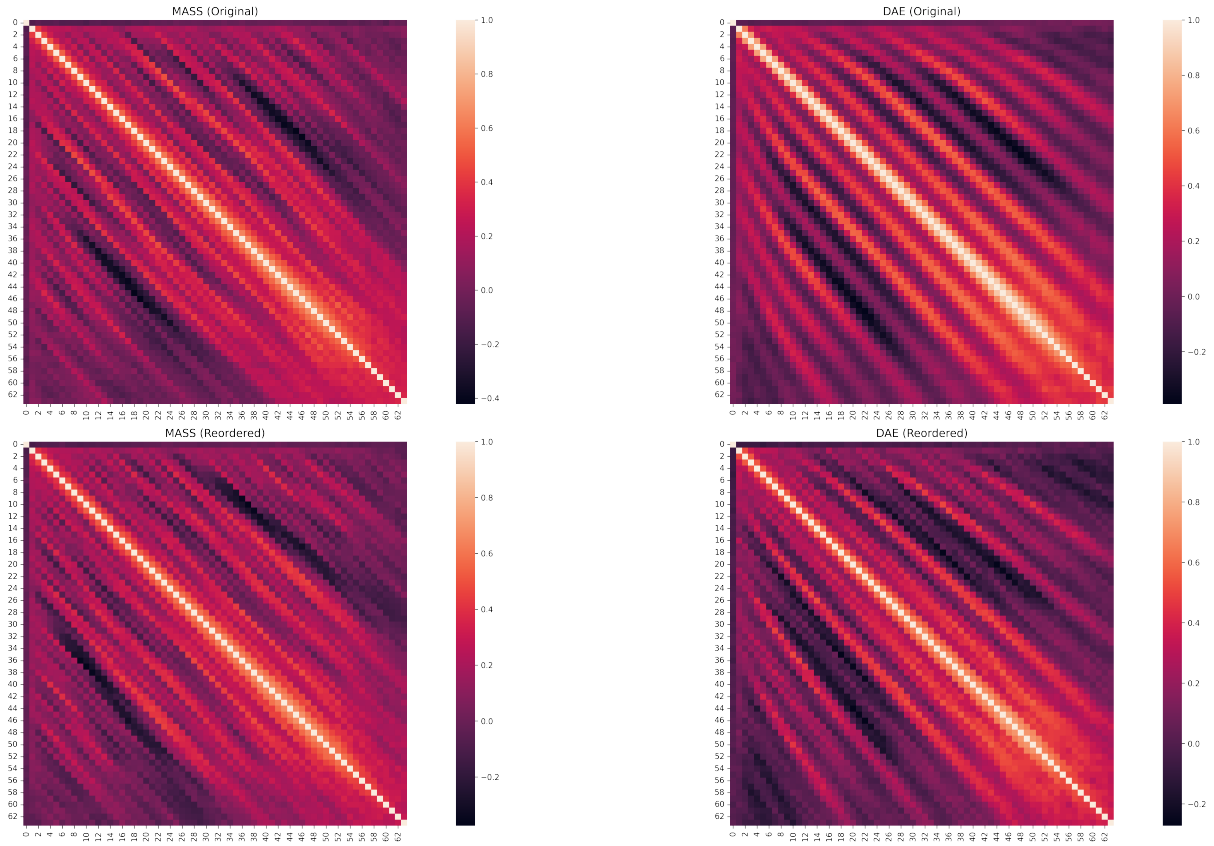


Figure 2: Comparison among the position embeddings of four UNMT models (*i.e.* MASS-original, DAE-original, MASS-reordered and DAE-reordered). Language pair: en→hi.

lations in the source language but in re-ordered form. The UNMT model trained on the DAE-static approach with original data achieved the highest performance for most language pairs.

Re-ordering effectively addresses word-order divergence in MASS-static models, revealing their sensitivity to this issue. The absence of shuffling in the pretraining objective limits the model's ability to retain position information, resulting in reduced performance. Unlike MASS-static models, DAE-static models perform well even without re-ordering, indicating their robustness to word-order divergence. Including shuffling in the DAE objective function allows the model to be more flexible in retaining or disregarding position information.

Re-ordering DAE-static models with re-ordered data surprisingly leads to a degradation in BLEU and CHRF scores. This is due to the ambiguity arising from re-ordering without case markers. An example illustrating this issue is presented in Figure 1, where the reference translation of the subject 'we' is 'hameM', which is in the dative case. However, when we re-order the English sentence, the subject remains 'we', which is in the subjective case and is frequently translated to 'hama' in Hindi. As a result, the re-ordered model incorrectly translates 'we' as 'hama' instead of the desired 'hameM'.

Figure 2 visualizes the position-wise cosine similarity of each position embedding. In MASS, position embeddings are similar to nearby positions within a local neighborhood of 2 or 3 positions. In DAE, position embeddings exhibit similarity to a larger local neighborhood, likely due to local shuffling noise. DAE models show more similarity to neighboring positions in the presence of word-order divergence (without re-ordering) compared to word-order similarity (with re-ordering).

## 6 Conclusion

Our findings show that DAE-based UNMT performs better than MASS-based UNMT in addressing word-order divergence. Furthermore, the DAE-based model is bidirectional, unlike the unidirectional re-ordered model. In future work, we aim to investigate additional word orders and language model objectives to gain deeper insights into the impact of word-order divergence in UNMT.

## References


Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642.

Tamali Banerjee, Rudra V Murthy, and Pushpak Bhattacharya. 2021. Crosslingual embeddings are essential in UNMT for distant languages: An English to IndoAryan case study. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 23–34, Virtual. Association for Machine Translation in the Americas.

Pushpak Bhattacharyya. 2012. Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilinguality. *CSI journal of computing*, 1(2):1–13.

Rajen Chatterjee, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2014. Supertag based pre-ordering in machine translation. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 30–38.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2021. Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online. Association for Computational Linguistics.

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

Jinhua Du and Andy Way. 2017. Pre-reordering for neural machine translation: Helpful or harmful? *The Prague bulletin of mathematical linguistics*, 108(1):171.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Jyotsana Khatri, Rudra Murthy, Tamali Banerjee, and Pushpak Bhattacharyya. 2021. Simple measures of bridging lexical divergence help unsupervised neural machine translation for low-resource languages. *Machine Translation*, 35(4):711–744.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, 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.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Rudra Murthy V, Anoop Kunchukuttan, Pushpak Bhattacharyya, 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*.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2016. chrf deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Haipeng Sun, Rui Wang, Masao Utiyama, Benjamin Marie, Kehai Chen, Eiichiro Sumita, and Tiejun Zhao. 2021. Unsupervised neural machine translation for similar and distant language pairs: An empirical study. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–17.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting pre-ordering for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).