# Unsupervised Neural Machine Translation

**Tutorial @ ICON-2020**
**IIT Patna**


Diptesh Kanojia
Jyotsana Khatri
Tamali Banerjee

**Prof. Pushpak Bhattacharyya**

Rudra Murthy






# Paradigms of Machine Translation

Pushpak Bhattacharyya

*Acknowledgement: Numerous PhD, masters and UG students and research staff working on MT with me since 2000*

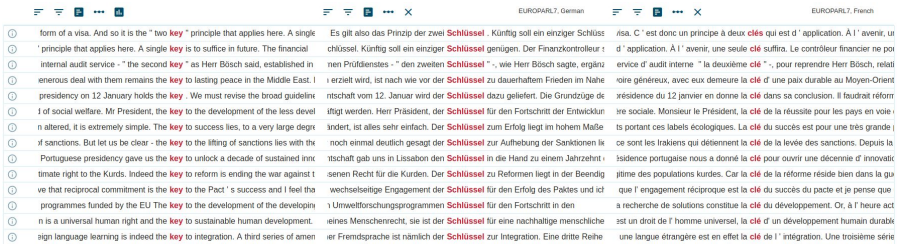
# Why is Unsupervised NMT needed?

**Diptesh Kanojia**

# Unsupervised NMT - Why?

## Supervised NMT

- Parallel Corpus
- Monolingual Corpus




Manual Translations ⟹ Cognitive Load

# “Unsupervised” NMT

- **No parallel corpus**

**However, the requirement is:**

- **Large monolingual corpus**
- Cross-lingual Word Embeddings
- **Low-resource languages**

Image Source: Paramount Pictures

# Resource Constraints

- Lack of resources for NLP tasks.

- Low resource languages.
  - Indian Languages including Sanskrit.
  - Hebrew, Greek, and Latin.

- Obscure Languages such as Sentinelese (North Sentinel Island, Indian Ocean), Ugaritic, etc.

- Monolingual corpus may be available.

# Resource Generation/Building

- Parallel word mappings can be generated.
  - Unsupervised Embedding mappings (similar script).

- Word mappings can also be created manually.
  - For language written in different scripts, but human supervision is needed.

- Word representations form the crux of most NLP tasks.

# Foundations

1. **Cross-lingual embeddings**
2. Denoising Autoencoder
3. Back-translation

# Word Representation for Humans

In humans, the acquisition of information and creation of mental representations occurs in a two-step process. (Ramos et. al., 2014)

> Sufficiently complex brain structure is necessary to establishing internal states capable to co-vary with external events.
>
> The validity or meaning of these representations must be gradually achieved by confronting them with the environment.

# Cross-lingual Word Embeddings

- The geometric relations that hold between words are similar across languages*.
    - For instance, numbers and animals in English show a similar (isomorphic) geometric structure as their Spanish counterparts.
- The vector space of a source languages can be transformed to the vector space of the target language t by learning a linear projection with a transformation matrix $W^{s \to t}$.

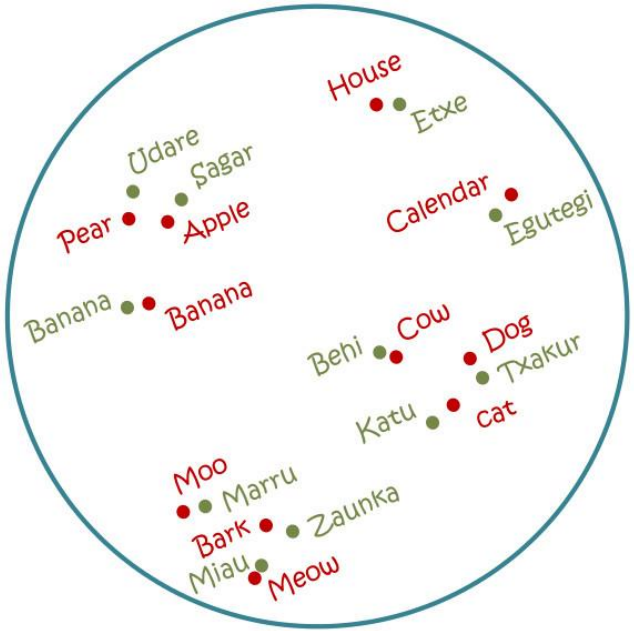
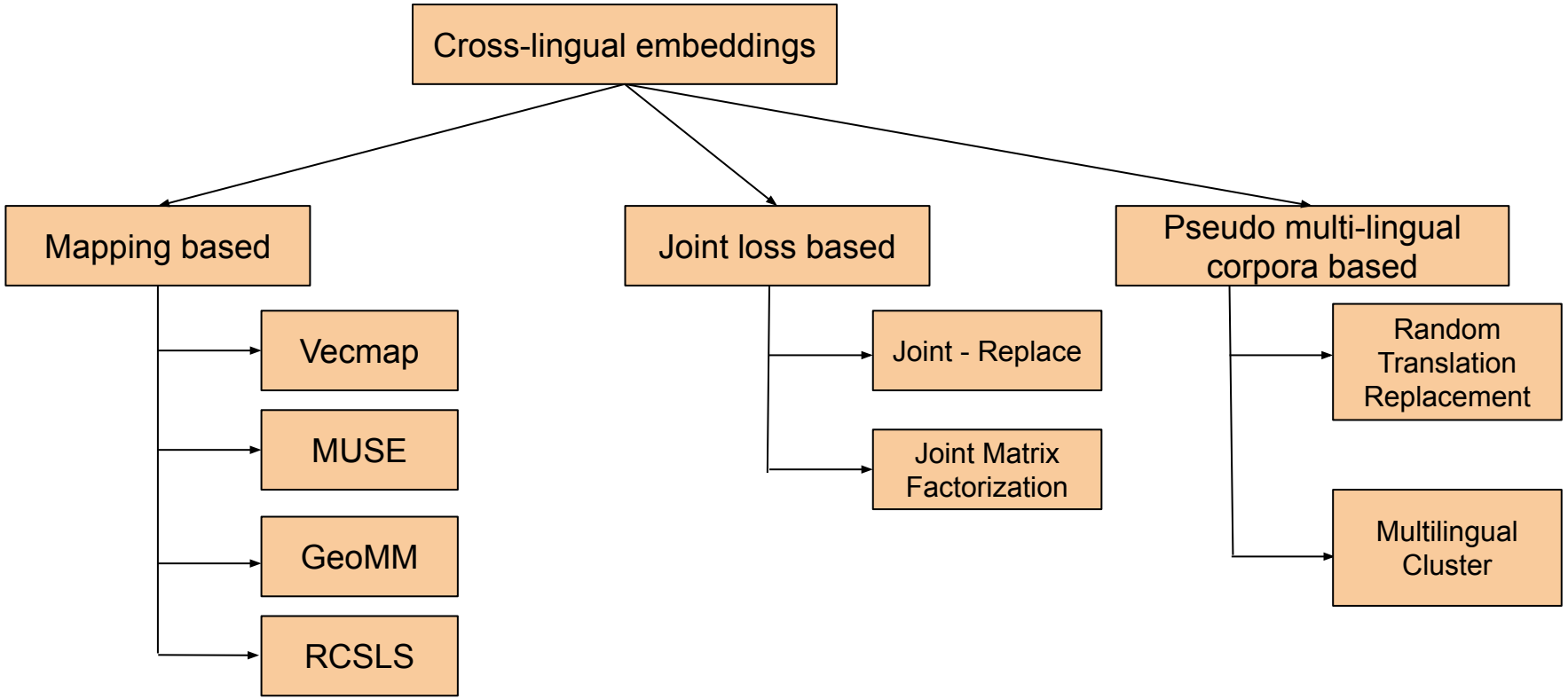
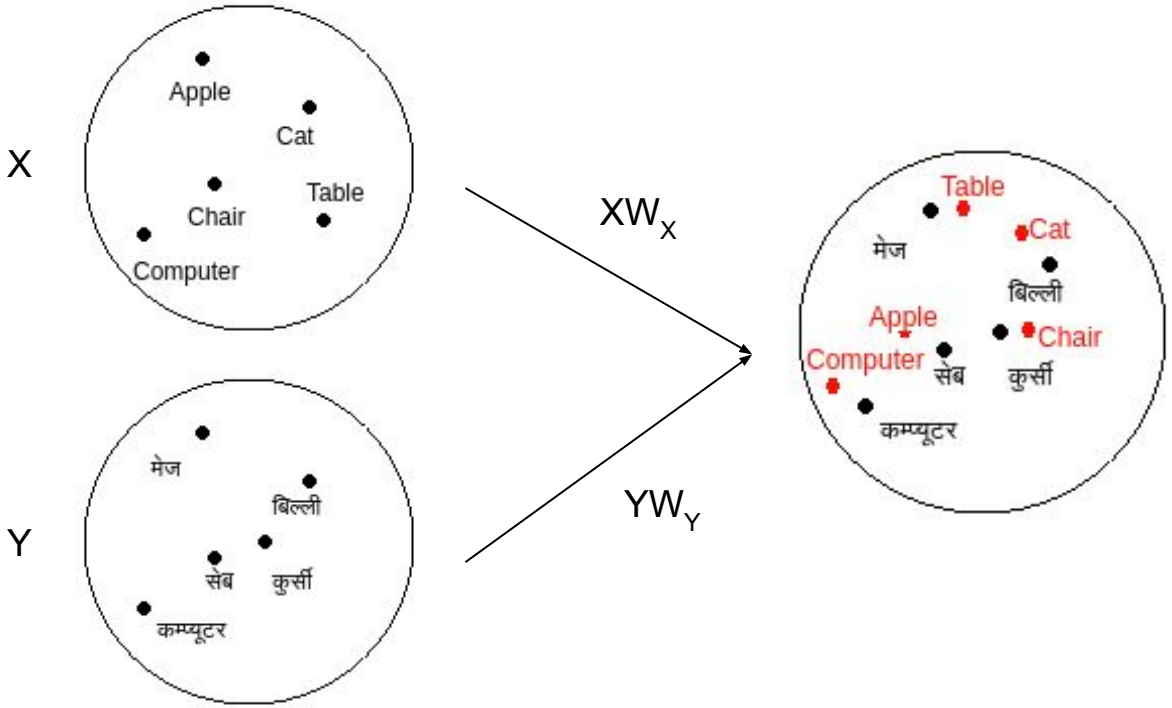



Image source- www.mikelartetxe.com

# Cross-lingual embeddings: Approaches



# Cross-lingual embeddings: Mapping based




- Task is to learn $\mathbf{W_X}$ and $\mathbf{W_Y}$ (the transformation matrices)
- X, Y are monolingual embedding spaces

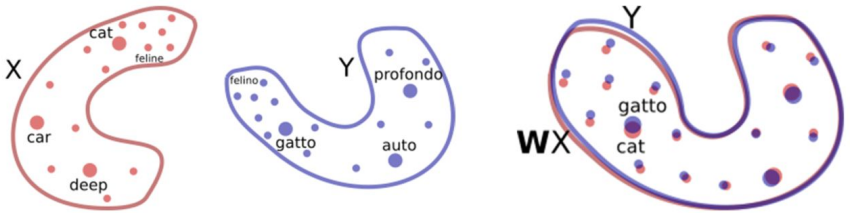
# MUSE

Given, target Vector Y and source Vector X

Learns Mapping Y=XW.

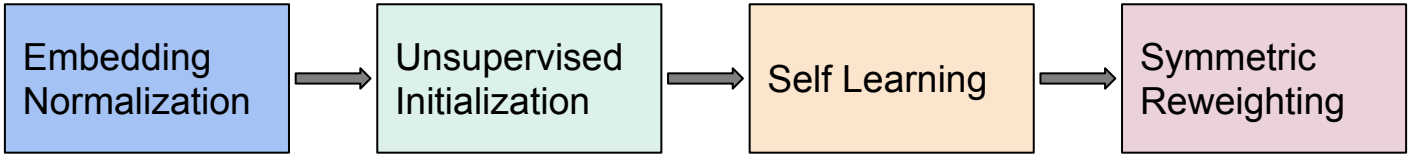
Trains a discriminator to tell whether two vectors are from the same language.

Also, a generator to map the vectors from one language into each other.




Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "Word translation without parallel data." *arXiv preprint arXiv:1710.04087* (2017).

# VecMap (Artexe et al. 2018)




- Embeddings Normalization
  - Length normalization + Mean centering + Length normalization
- Unsupervised initialization
  - Assume both spaces are isometric
  - Nearest neighbor retrieval on $XX^T$ and $YY^T$
- Self training
  - Compute the optimal orthogonal mapping by maximizing the similarity for the current dictionary D
  - Compute the dictionary over the similarity matrix of the mapped embeddings
- Symmetric weighting to induce good dictionary
  - $W_X = US^{1/2}$ , $W_Y = VS^{1/2}$

Artetxe Mikel, Gorka Labaka, and Eneko Agirre. "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings." *ACL 2018*.

# Joint training + Cross-lingual alignment (Wang et al 2019)

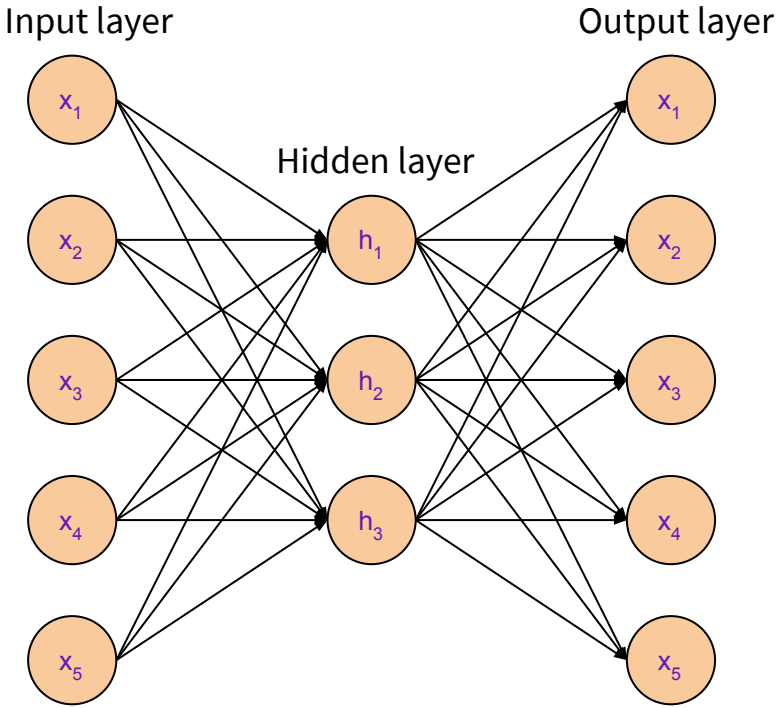
- Joint initialization
  - Joint training using monolingual embedding training algorithm using combined corpus
- Vocabulary reallocation
  - Create source, target and common vocabulary
- Alignment refinement
  - Mapping based algorithm for align source and target to the same space

Wang Z, Xie J, Xu R, Yang Y, Neubig G, Carbonell JG (2019) Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In: International Conference on Learning Representations

# Foundations

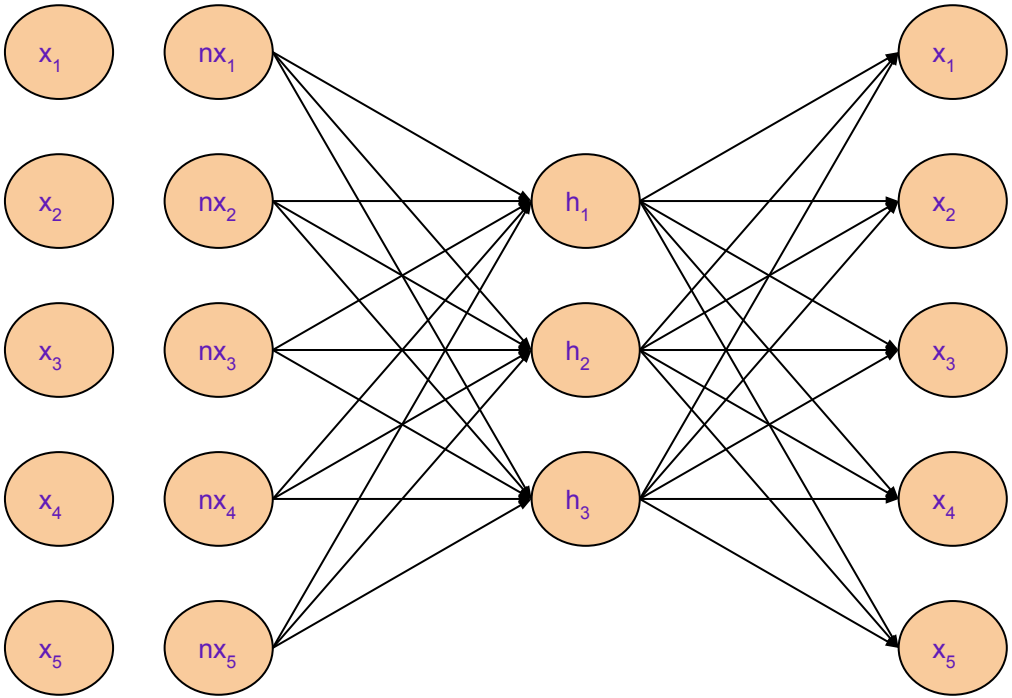
1. Cross-lingual embeddings
2. **Denoising Autoencoder**
3. Back-translation

# Autoencoder




- Representation learning
- Neural network to learn reconstruction of the data
- Optimize **Reconstruction Error**
- Balance between
  - Accurately build a reconstruction
  - Handle inputs such that the model doesn't learn to copy the data

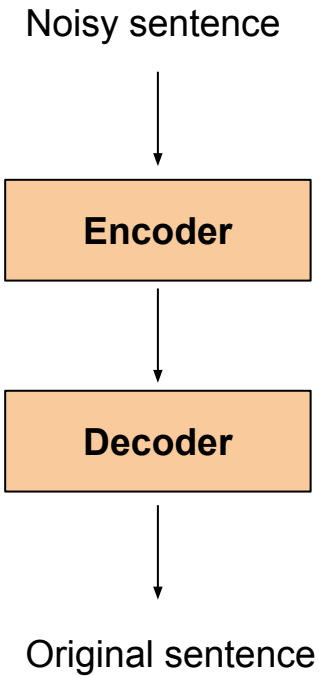
# Denoising auto-encoder




Corrupted data

- Learn to generate original sentence from a noisy version of it
- Eliminates the learning of identity function

# Denoising auto-encoder



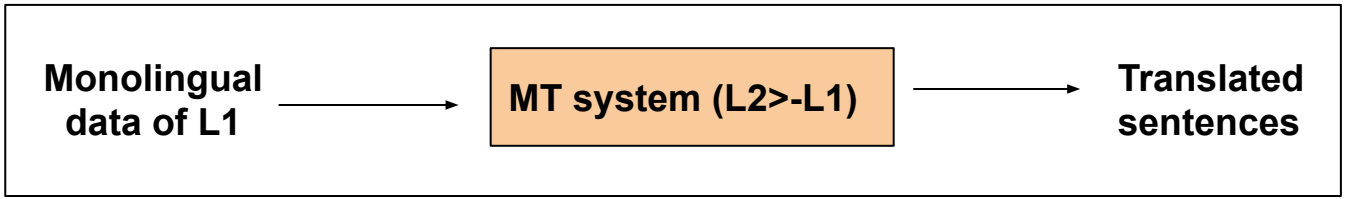

- Encoder representation is the representation for noisy sentence
- Decoder tries to generate the original sentence from the encoder representation of the noisy sentence
- A sentence can be corrupted using different types of noise
  - Swapping of words
  - Removal of words
  - Replacement of words with other words

# Foundations

1. Cross-lingual embeddings
2. Denoising Autoencoder
3. **Back-translation**

# Back-Translation

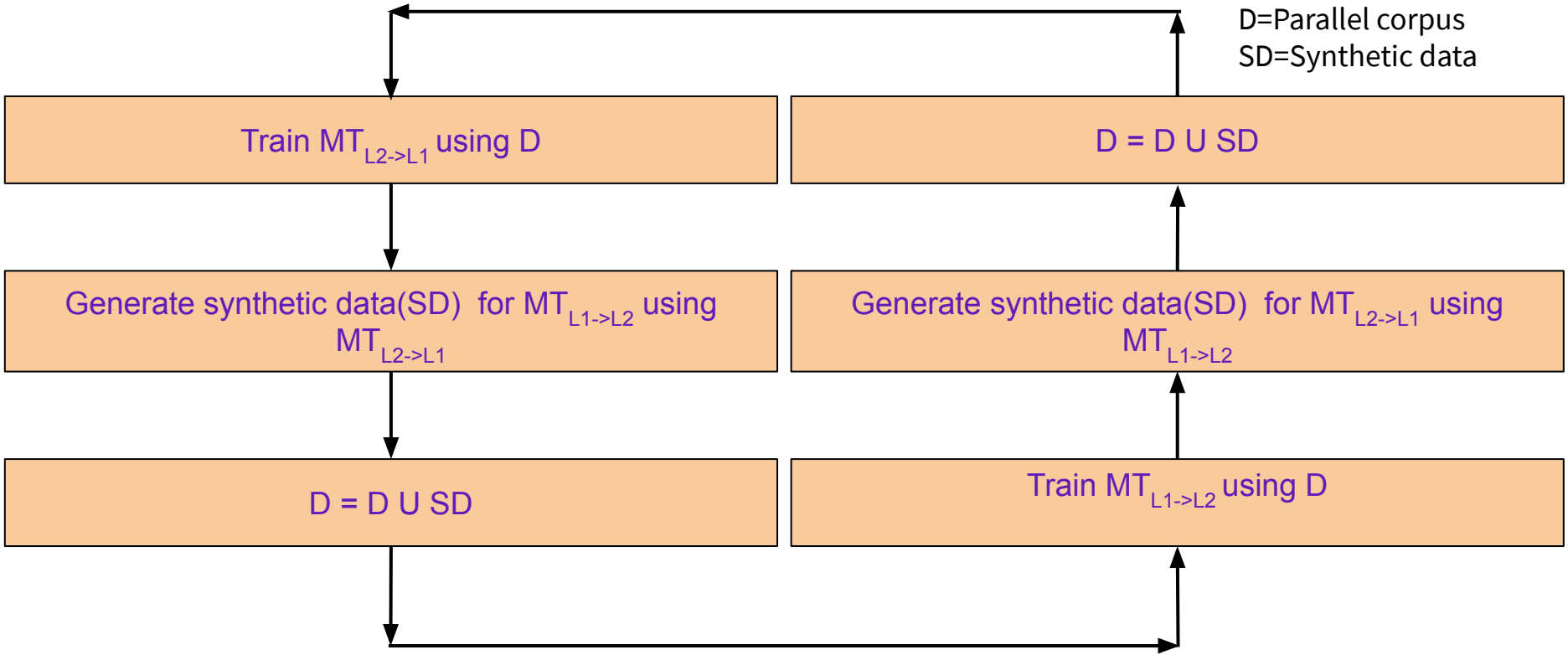
- Utilize monolingual data of target language
- Generate pseudo parallel data using MT system in opposite direction (target->source)




- Train MT system (L1->L2) using a combination of parallel and generated synthetic data both

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Improving Neural Machine Translation Models with Monolingual Data." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96. 2016.

# Iterative Back-Translation



# Iterative Back-Translation

| Setting | French–English | | English–French | | Farsi–English | English-Farsi |
|---|---|---|---|---|---|---|
| | 100K | 1M | 100K | 1M | 100K | 100K |
| NMT baseline | 16.7 | 24.7 | 18.0 | 25.6 | 21.7 | 16.4 |
| back-translation | 22.1 | 27.8 | 21.5 | 27.0 | 22.1 | 16.7 |
| back-translation iterative+1 | 22.5 | - | 22.7 | - | 22.7 | 17.1 |
| back-translation iterative+2 | 22.6 | - | 22.6 | - | 22.6 | 17.2 |

- Beneficial for Low resource languages also

Image source: Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. "Iterative back-translation for neural machine translation." In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24. 2018.

# UMT Approaches

Tamali Banerjee

1. **Unsupervised NMT**
2. GAN for UNMT
3. Unsupervised SMT
4. Hybrid UMT

# Introduction

- In ICLR 2018, two concurrent papers showed that it is possible to train an NMT system without using any parallel data.

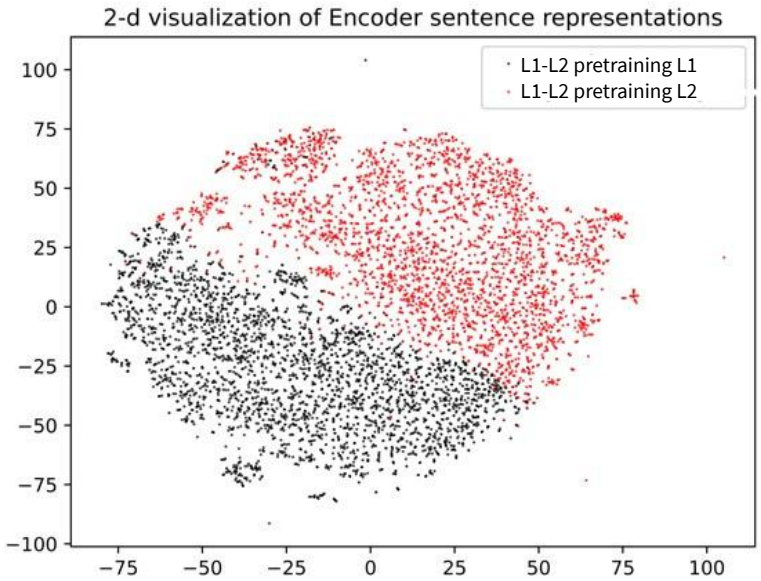
List of papers


1. Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).
2. G. Lample, A. Conneau, L. Denoyer, MA. Ranzato. 2018. Unsupervised Machine Translation With Monolingual Data Only. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

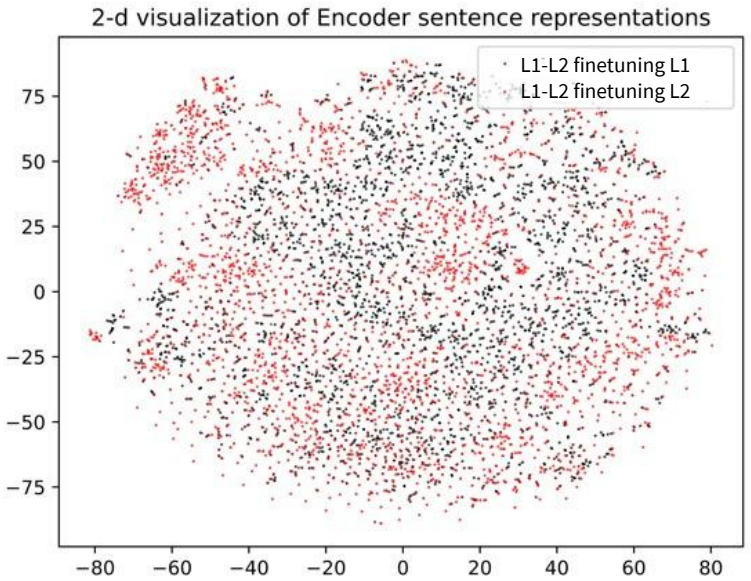
# Components of U-NMT

- Bi-lingual embedding: It projects word embeddings of both languages in the same embedding space.
- Language modeling: It helps the model to encode and generate sentences.
    - Through initialization of the translation models.
    - Through iterative training.
- Iterative  back-translation: It bridges the gap between encoder sentence representation in source and target languages.

# Effect of Back-translation




Before Back-translation




After Back-translation

Image credit: Rudra and Jyotsana

# Architecture

- Bi-lingual embedding layer
- Encoder-Decoder architecture
- Dual structure
- Sharing of modules

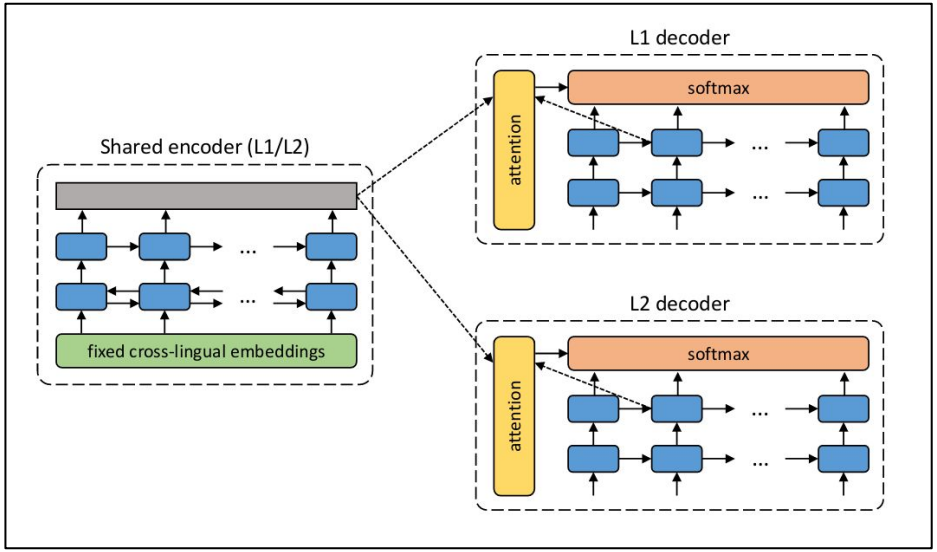
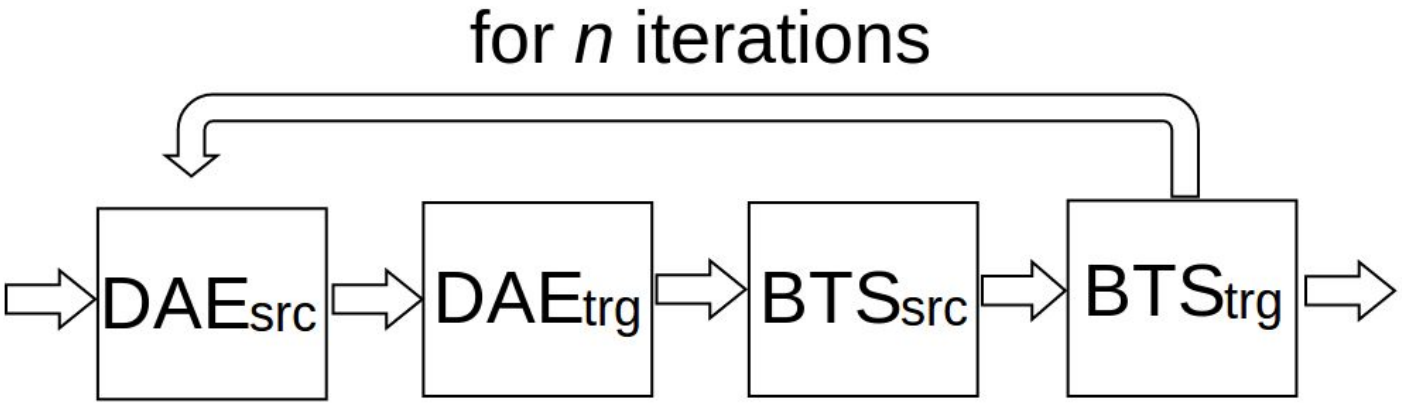


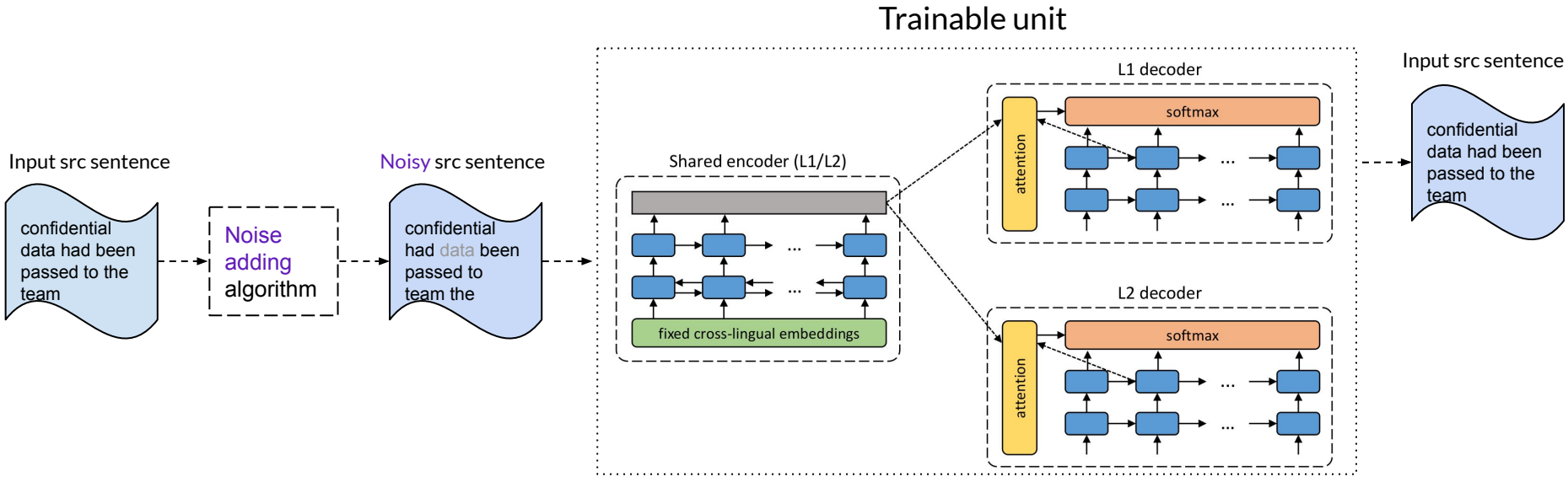

Image source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# Training Procedure



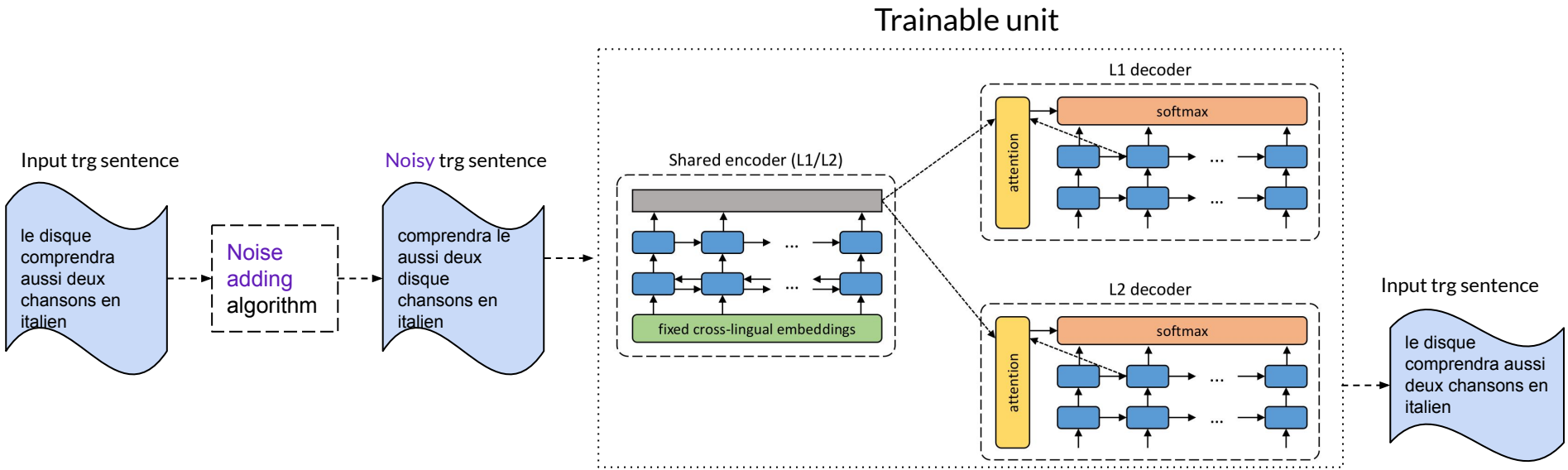

DAE$_{src}$: Denoising of source sentences; DAE$_{trg}$: Denoising of target sentences;
BTS$_{src}$: Back-translation with shuffled source sentences; BTS$_{trg}$: Back-translation with shuffled target sentences;
*n* : total number of iteration till it reaches stopping criterion.

# U-NMT: Denoising of source sentences



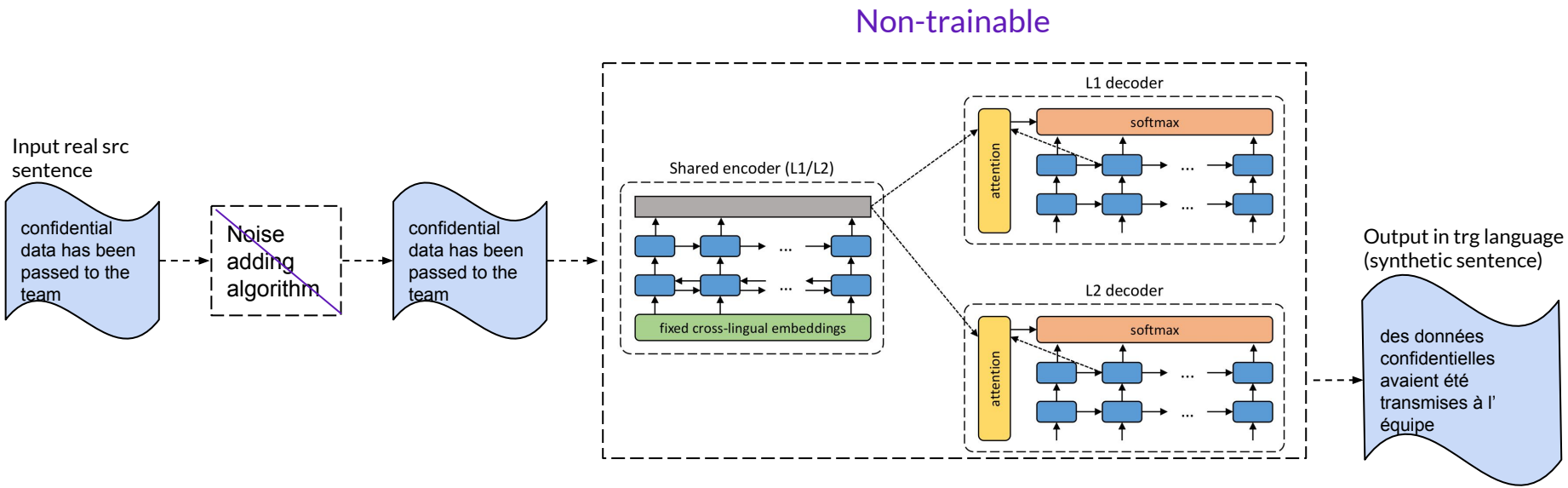

Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# U-NMT: Denoising of target sentences



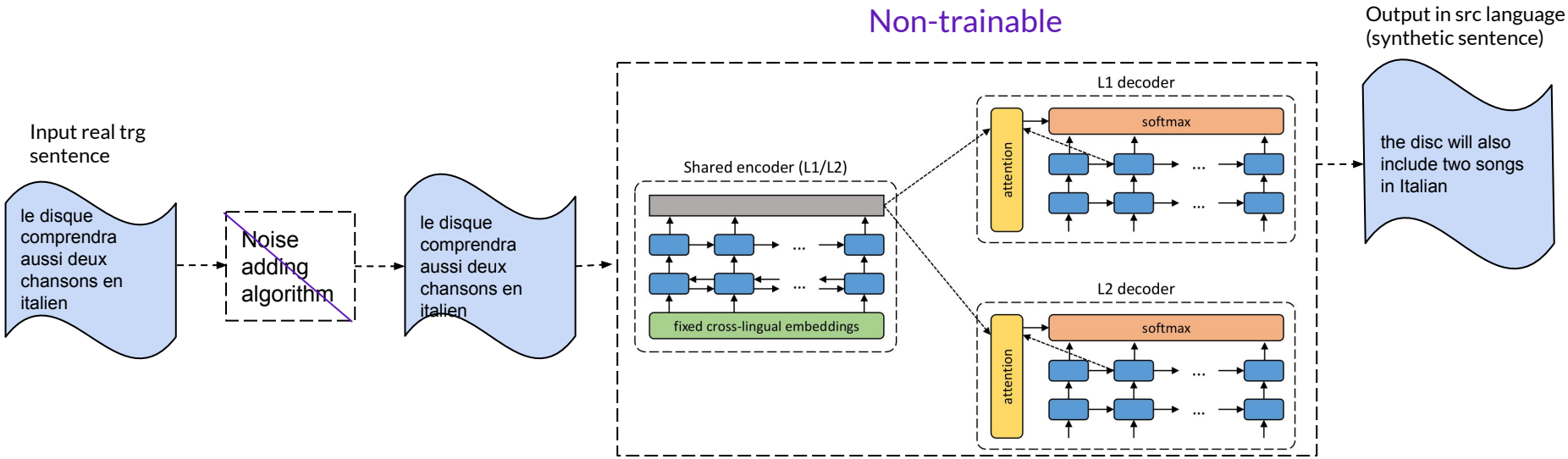

Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# U-NMT: Back-translation Corpus Construction (source to target)



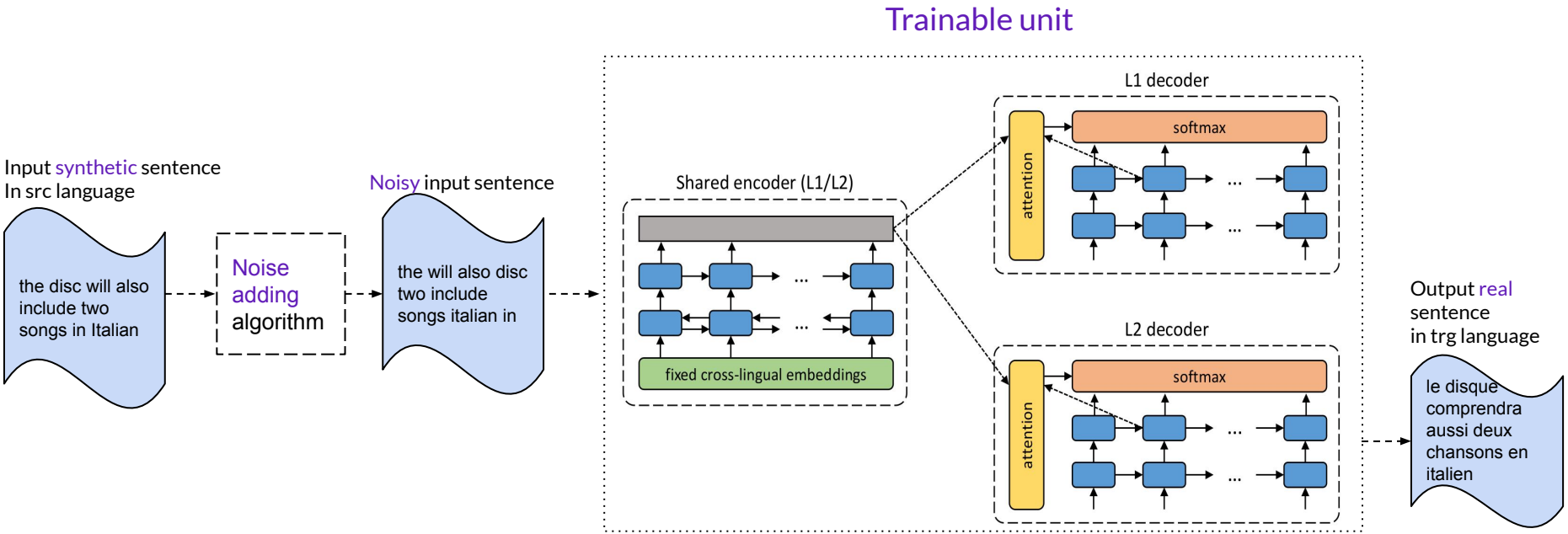

Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# U-NMT: Back-translation Corpus Construction (target to source)



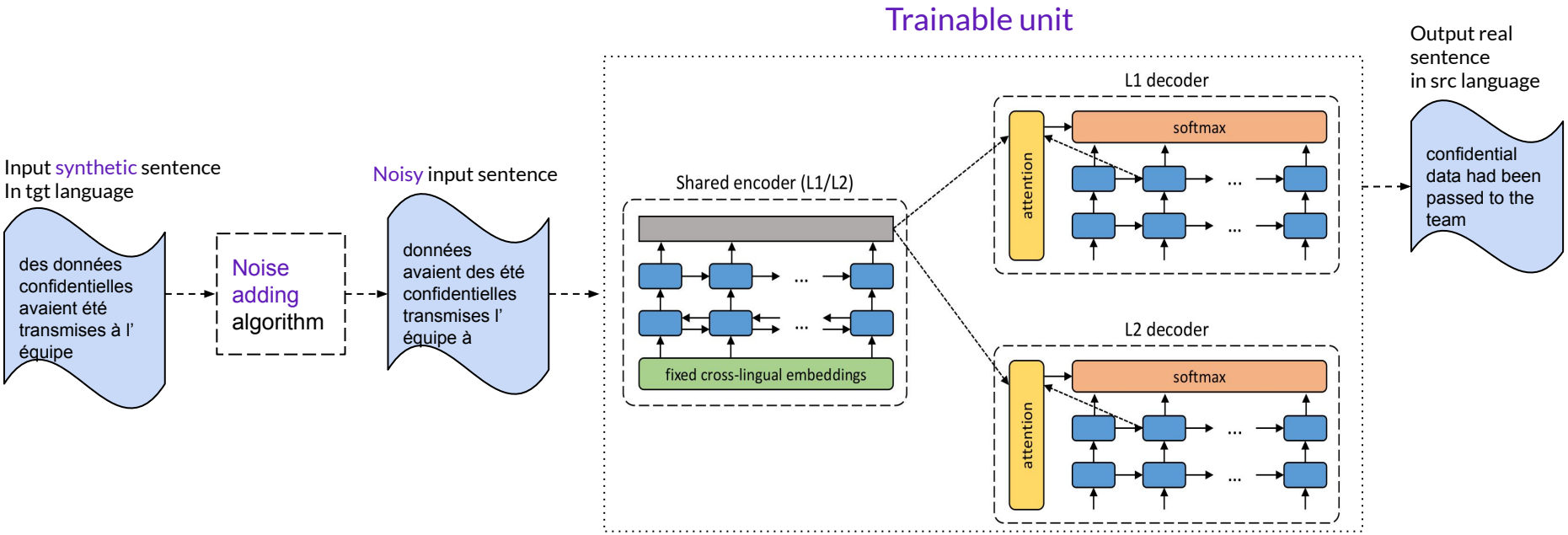

Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# U-NMT: Training with Back-translated data (source to target)



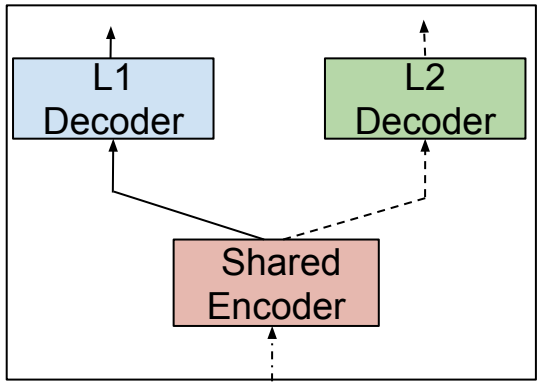

Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# U-NMT: Training with Back-translated data (target to source)

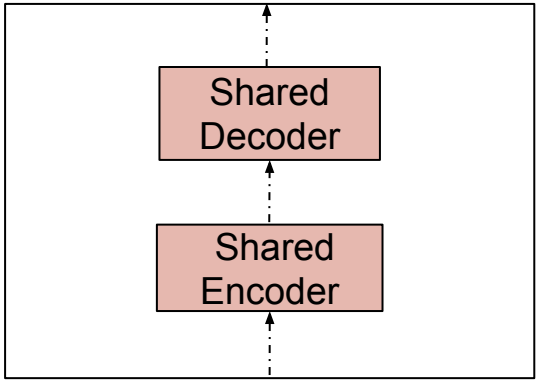



Source: Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

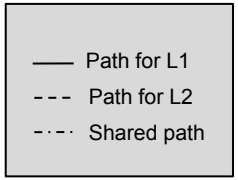
# Comparison between two approaches




Artexte et al.

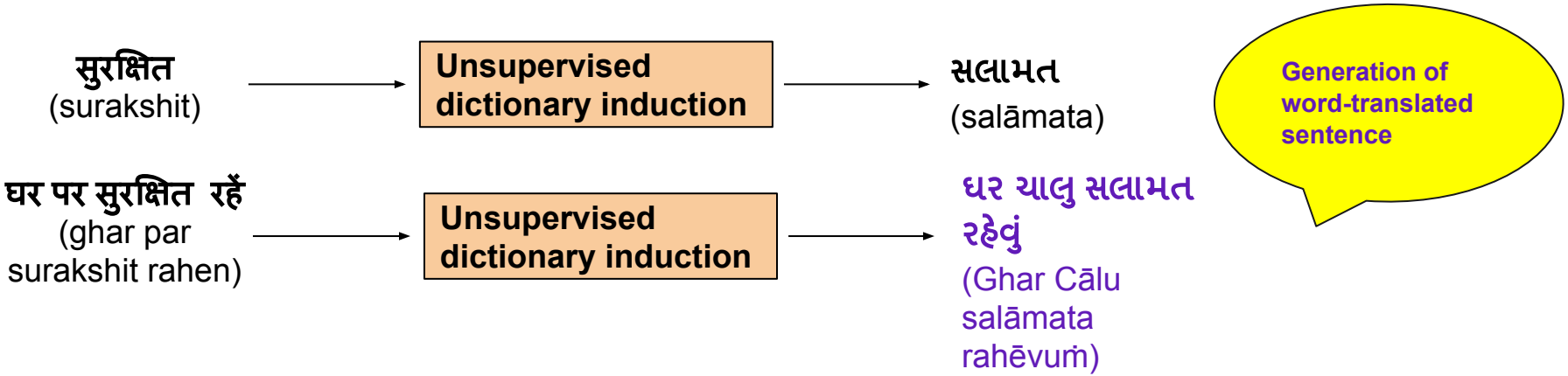



Lample et al.

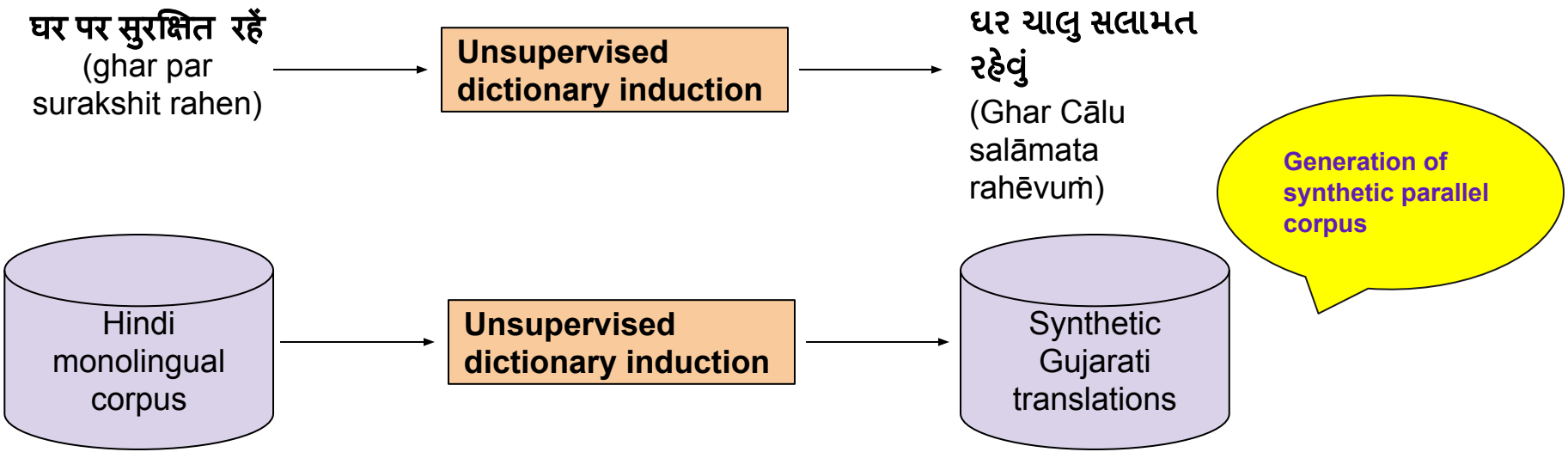



- Decoders are non-shared for Artexte et al. and shared for Lample et al.
- Lample et al. initialises training with word-by-word translation. [Next few slides]
- Lample et al. uses a language discriminator for encoder representation. It challenges the language invariance nature of encoder representations. [Next subsection]

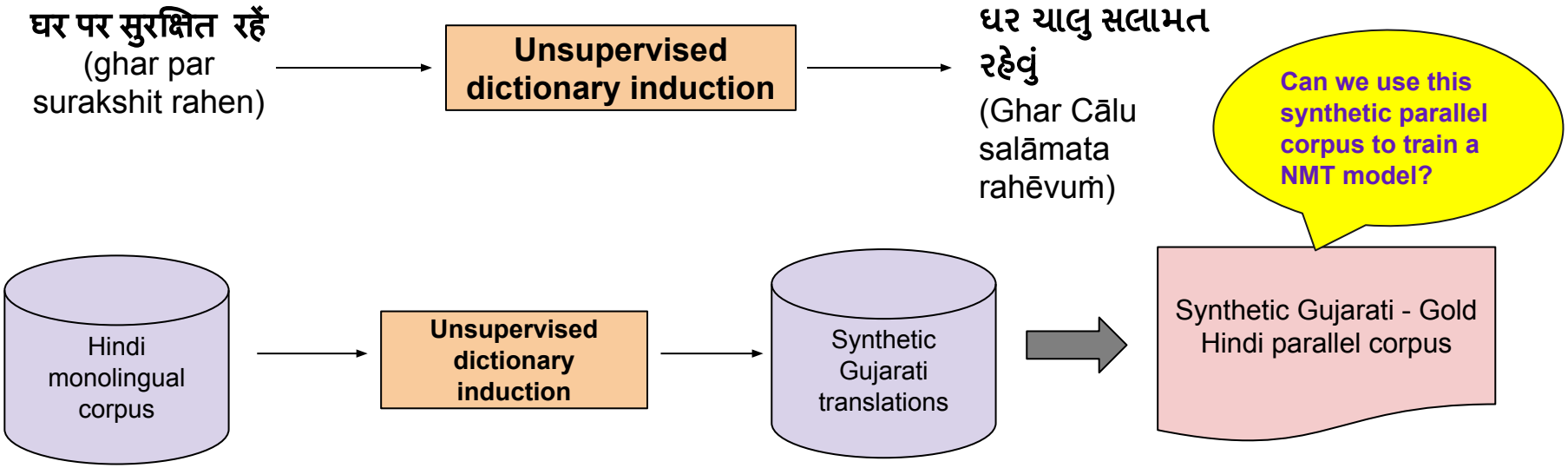
# Training with word-by-word translation



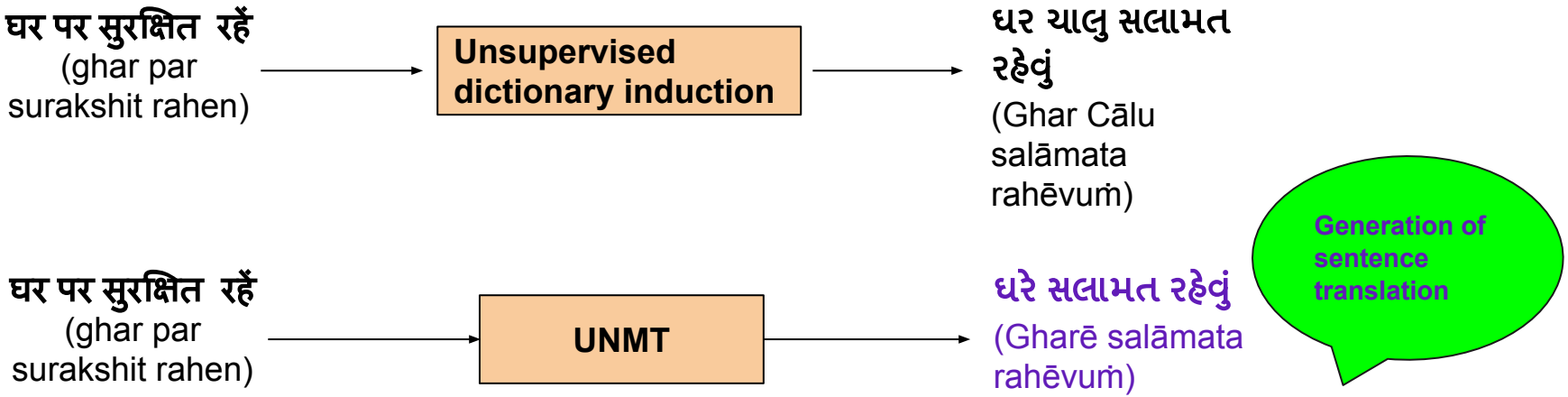
# Training with word-by-word translation



# Training with word-by-word translation



# Training with word-by-word translation



# Effect of DAE and BT

| Author | Approach | Fr → En | En→ Fr | De → En | En → De |
|---|---|---|---|---|---|
| Artexte et al. (tested on WMT14) | Emb. nearest neighbour | 9.98 | 6.25 | 7.07 | 4.39 |
| | Denoising | 7.28 | 5.33 | 3.64 | 2.40 |
| | Denoising + Back-translation | **15.56** | **15.13** | **10.21** | **6.55** |
| Lample et al. (tested on WMT14 en-fr and WMT16 en-de) | Emb. nearest neighbour | 10.09 | 6.28 | 10.77 | 7.06 |
| | Word2word pretraining + Denoising + Back-translation | **15.31** | **15.05** | **13.33** | **9.64** |

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

G. Lample, A. Conneau, L. Denoyer, MA. Ranzato. 2018. Unsupervised Machine Translation With Monolingual Data Only. In Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018).

# UMT Approaches

1. Unsupervised NMT
2. **GAN for UNMT**
3. Unsupervised SMT
4. Hybrid UMT

# Introduction

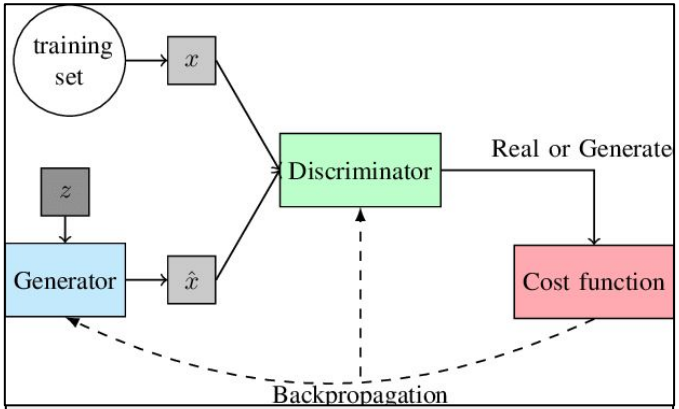
- Use GAN to enhance the language invariance.
- Sharing of the whole model faces difficulty in keeping the diversity of languages.
    - Share module partially

List of papers

1. Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

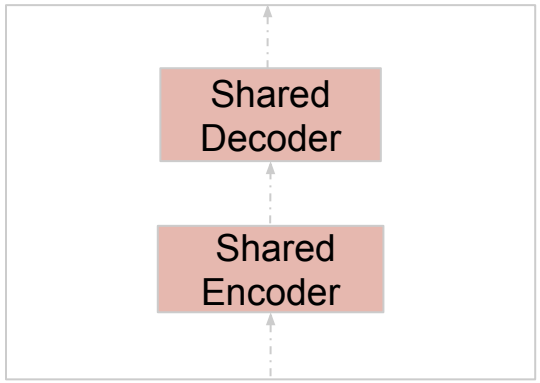
# Generative Adversarial Networks (GAN)

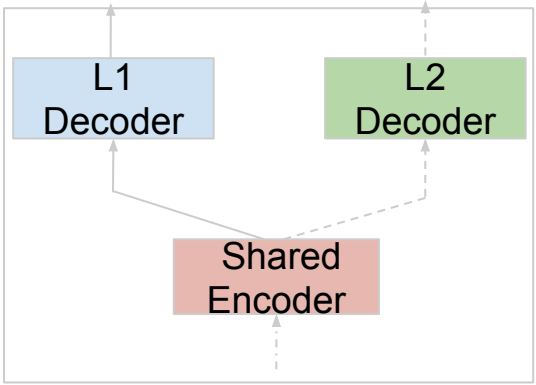
- GANs are a clever way of training with two sub-models:
  - Generator model that we train to generate new examples,
  - Discriminator model that tries to classify examples as either real.
- In case of UNMT,
  - Shared encoder is the generator.
  - An extra discriminator module is attached with it to discriminate encoder representations w.r.t. language.

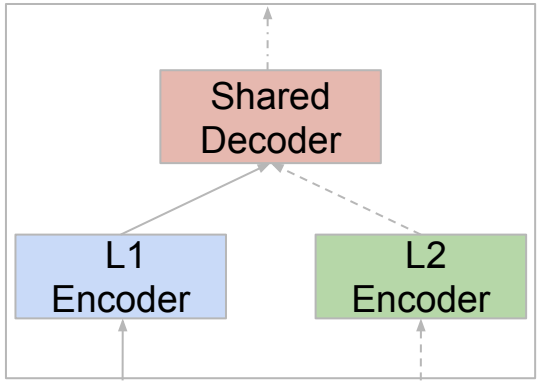


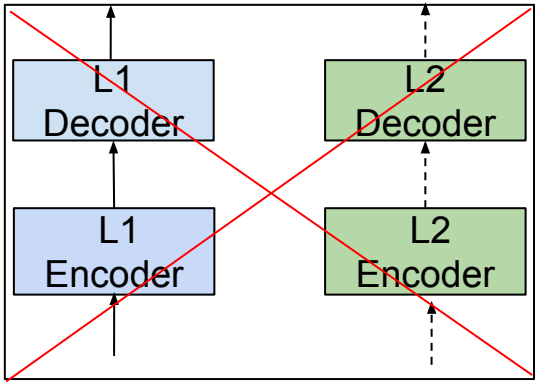

- GAN: Two neural networks (a generative network and a discriminative network) compete with each other to become more accurate in their predictions.

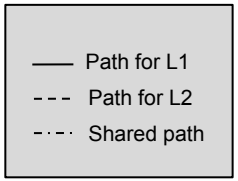
# Different parameter sharing strategies



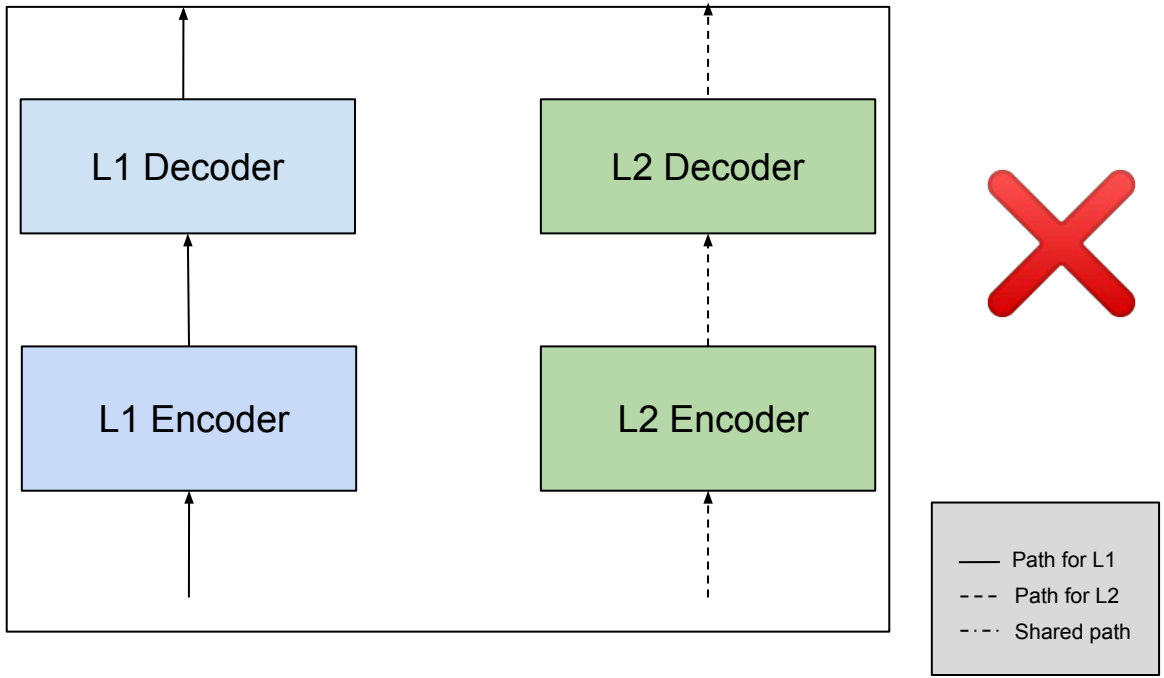



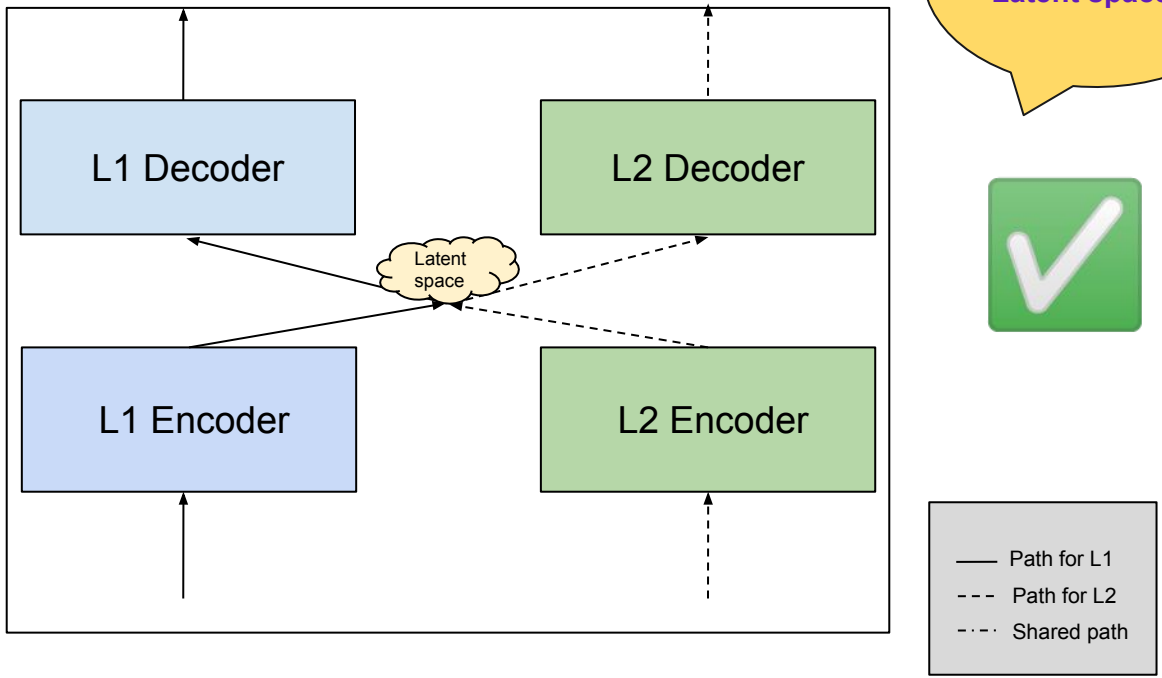



# Language specific Encoder-Decoder

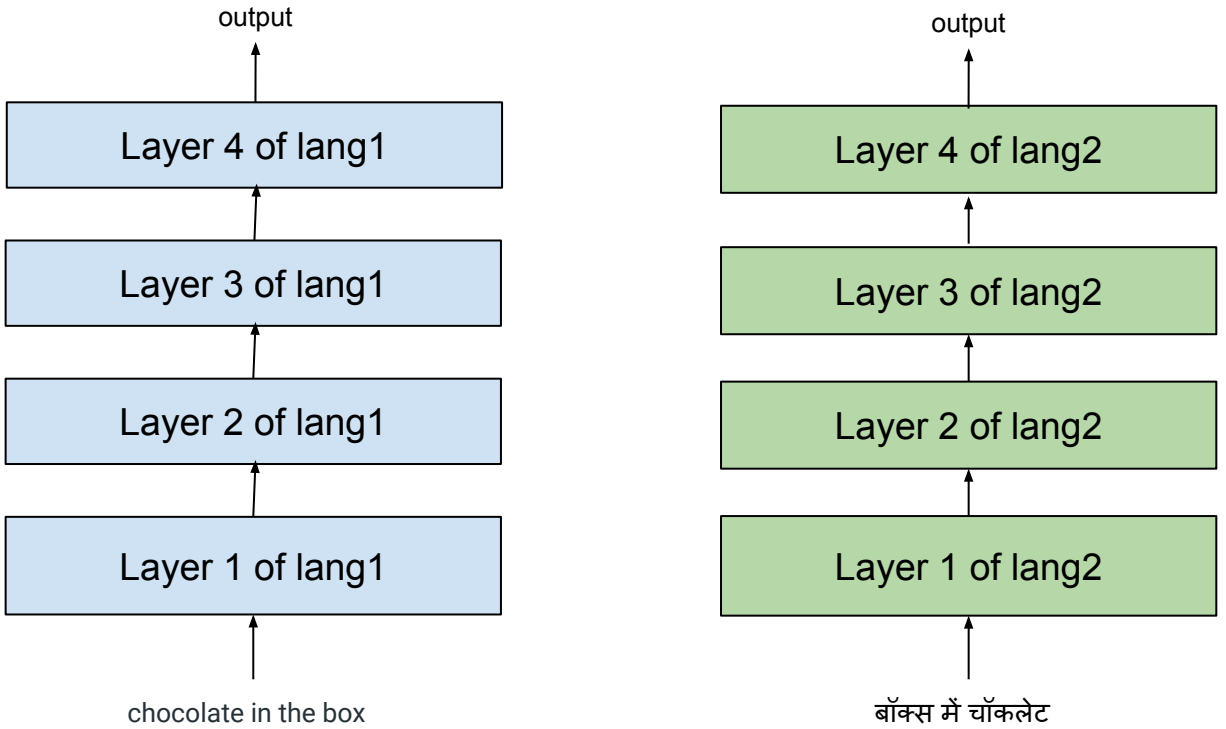


# Language specific Encoder-Decoder

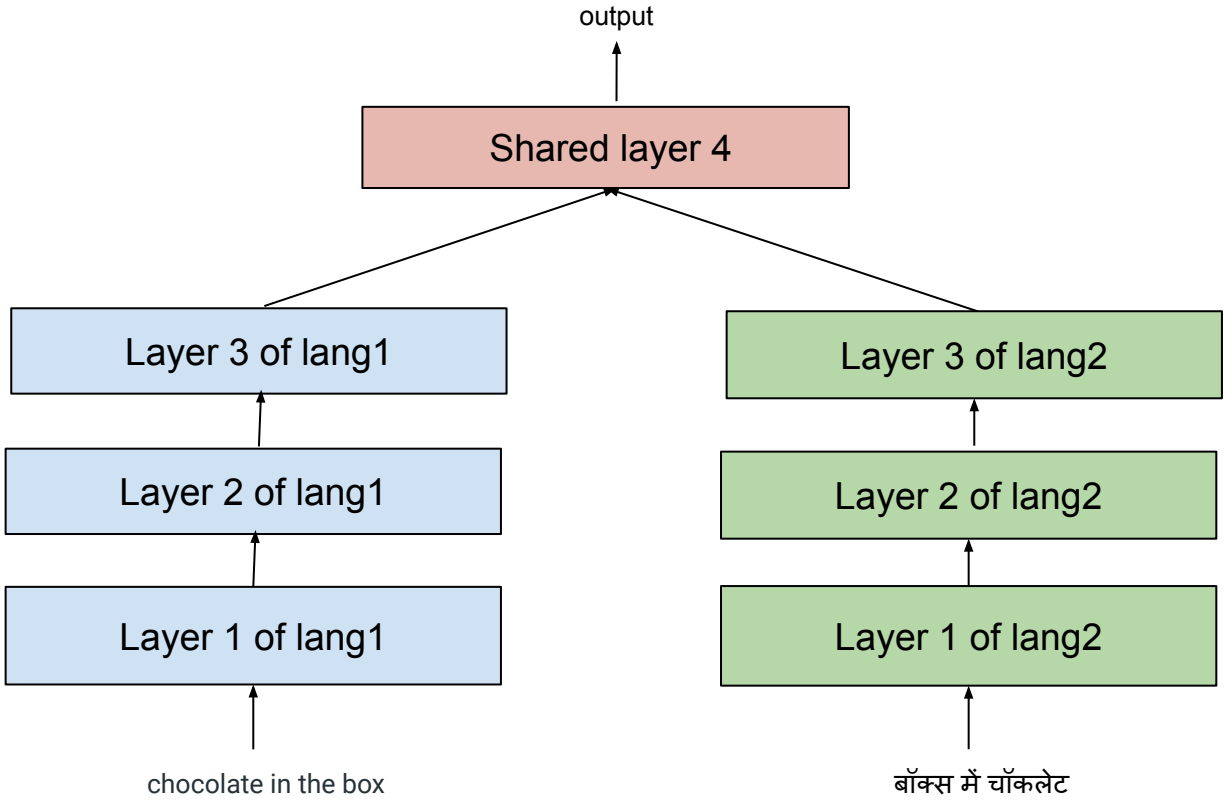



How to share Latent space?

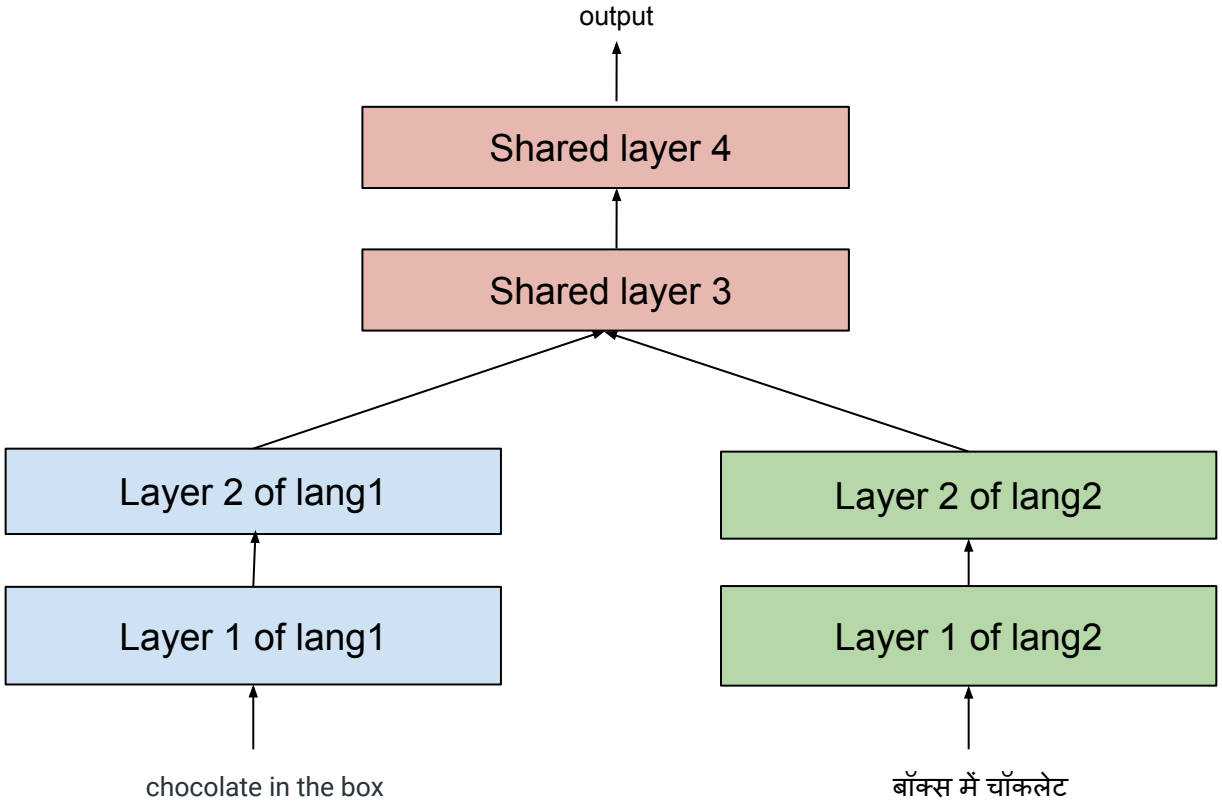
# Parameter sharing



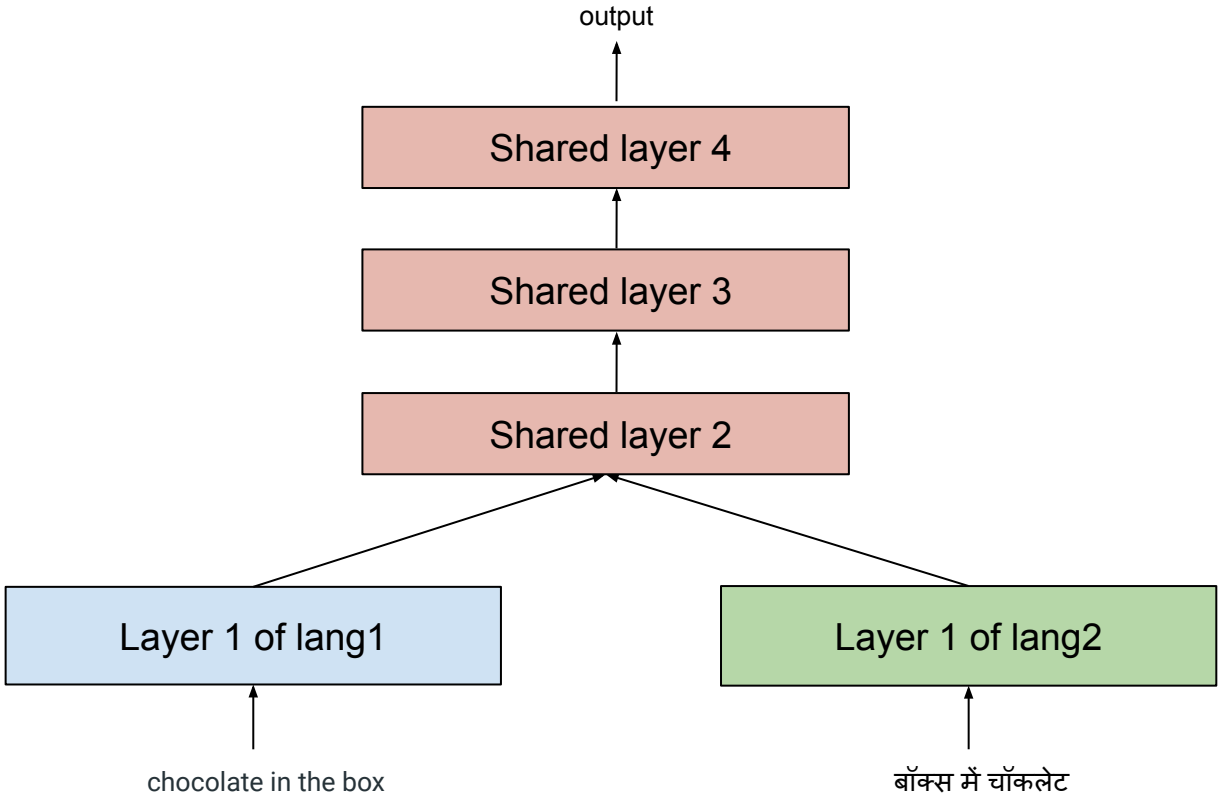
# Parameter sharing



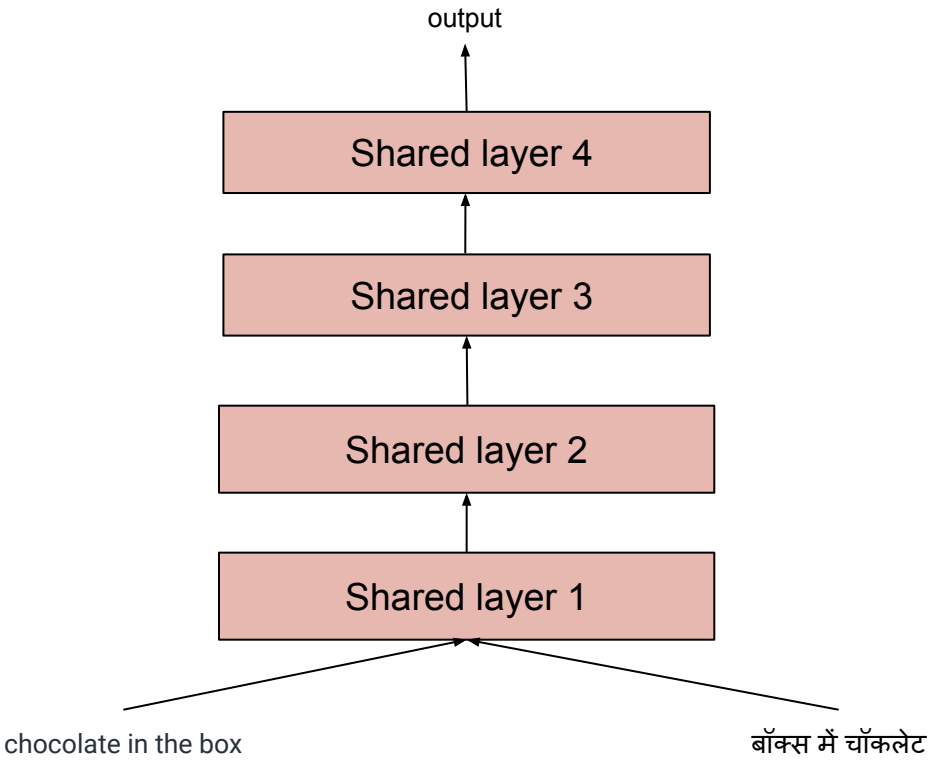
# Parameter sharing



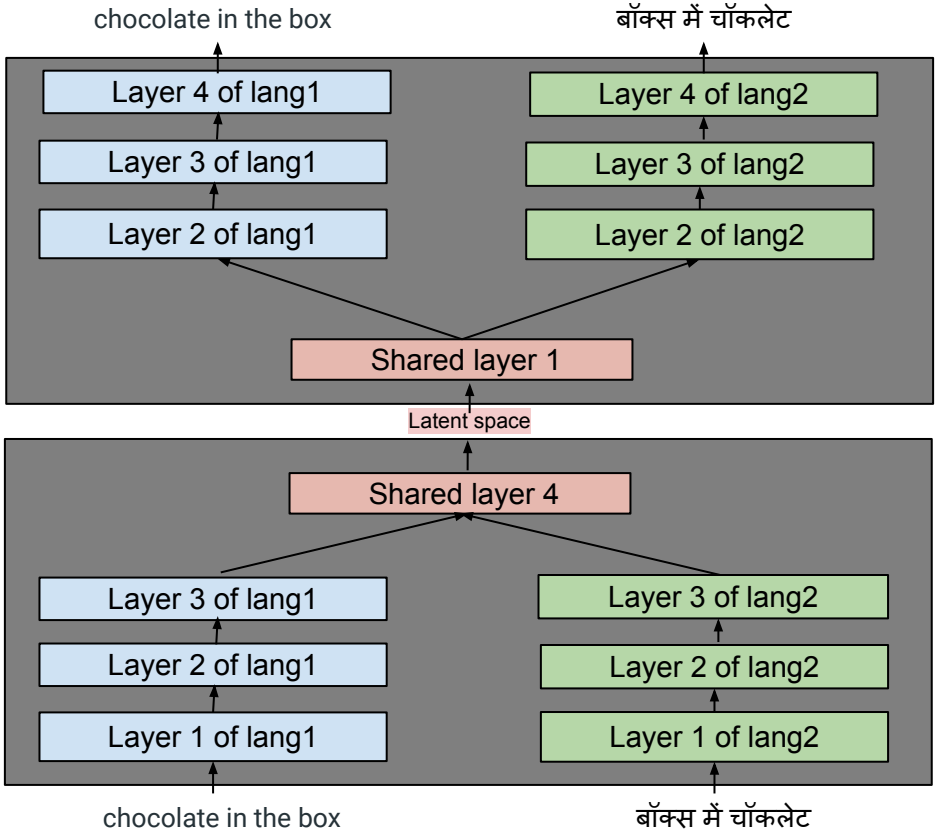
# Parameter sharing



# Parameter sharing



# Architecture with weight-sharing layers



# Number of weight-sharing layers vs. BLEU

- In this approach, sharing only 1 layer gives best BLEU scores.
- When sharing is more than 1 layer, the BLEU scores drop.
- This drop is more in case of distant language-pairs when compared to drop in close language-pairs.

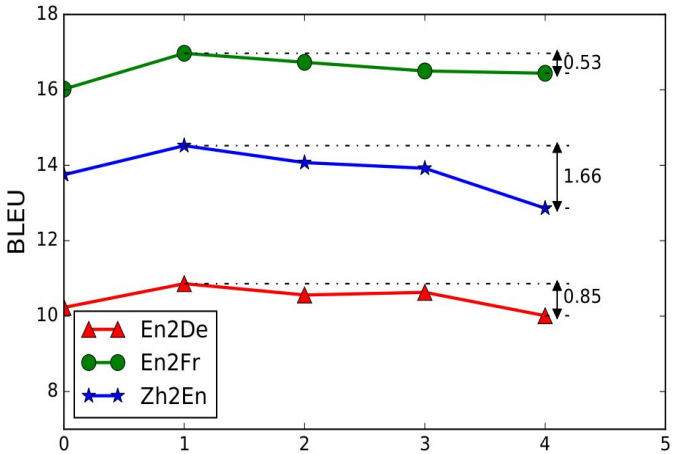
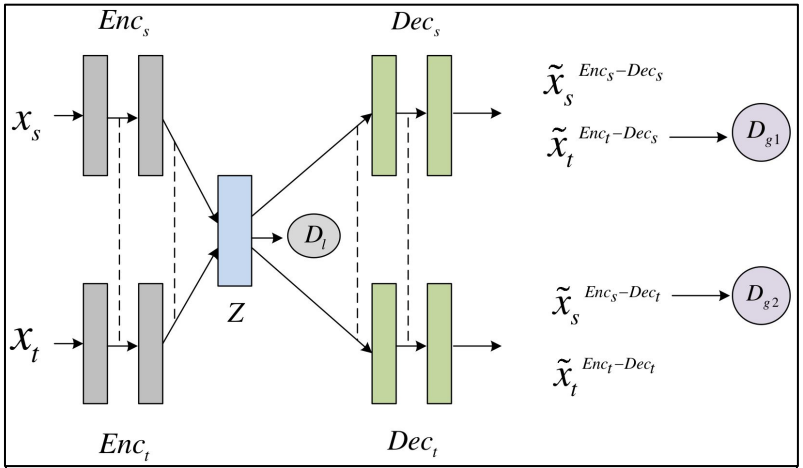



Image source: Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

# Weight sharing in UNMT

- When sharing is less, we need GAN to ensure input language invariance of encoder representations and outputs.
- Two types of GAN are used here.
  - Local GAN $D_L$ to ensure input language invariance of encoder representations.
  - Global GAN $D_{g1}$ and $D_{g2}$ to ensure input language invariance of output sentences.




Weight sharing UNMT architecture with GAN

Image source: Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

# Results

| | en-de | de-en | en-fr | fr-en | zh-en |
|---|---|---|---|---|---|
| Supervised | 24.07 | 26.99 | 30.50 | 30.21 | 40.02 |
| Word-by-word | 5.85 | 9.34 | 3.60 | 6.80 | 5.09 |
| Lample et al. (2017) | 9.64 | 13.33 | 15.05 | 14.31 | - |
| **The proposed approach** | **10.86** | **14.62** | **16.97** | **15.58** | **14.52** |

Yang, Z., Chen, W., Wang, F. and Xu, B., 2018, July. Unsupervised Neural Machine Translation with Weight Sharing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 46-55).

# UMT Approaches

1. Unsupervised NMT
2. GAN for UNMT
3. **Unsupervised SMT**
4. Hybrid UMT

# Introduction

- Components of SMT:
  1) Phrase table
  2) Language model
  3) Reordering model
  4) Word/phrase penalty
  5) Tuning
- Challenges-
  - Phrase table induction without parallel data.
  - Unsupervised Tuning
- Improvement-
  - Iterative refinement
  - Subword information

## List of papers


1. Artetxe, M., Labaka, G. and Agirre, E., 2018. Unsupervised Statistical Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3632-3642).

2. Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.A., 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049).

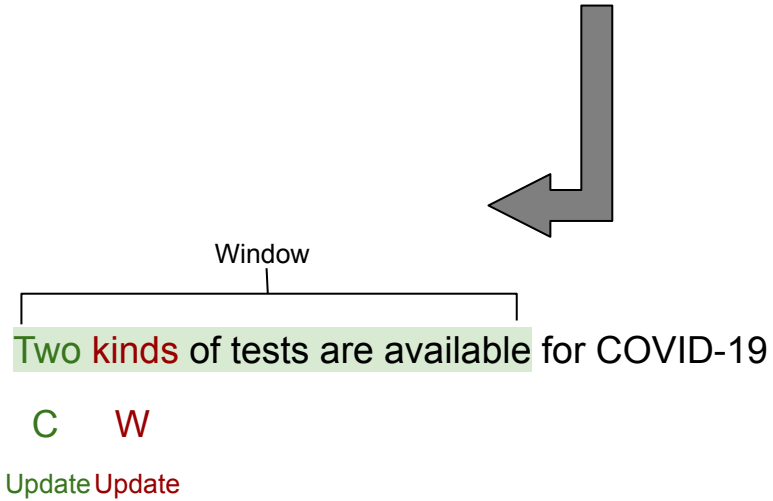
3. Artetxe, M., Labaka, G. and Agirre, E., 2019, July. An Effective Approach to Unsupervised Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 194-203).

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.

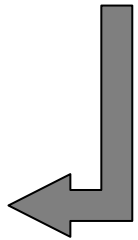
# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



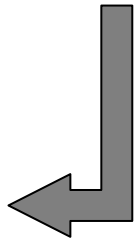
Two kinds of tests are available for COVID-19

W C

Update Update

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



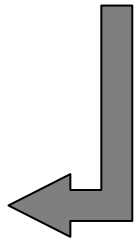
Two kinds of tests are available for COVID-19

W C

Update Update

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



Two kinds of tests are available for COVID-19

W

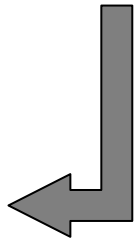
Update

C

Update

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



Two kinds of tests are available for COVID-19

W

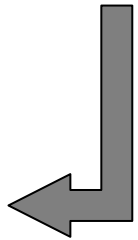
Update

C

Update

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



Two kinds of tests are available for COVID-19

P

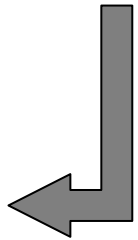
Update

C

~~Update~~

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



Two kinds of tests are available for COVID-19

P

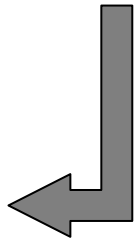
Update

C

~~Update~~

# Phrase table induction in an unsupervised way

- Get n-gram embedding using skip-gram with negative samples.



Two kinds of tests are available for COVID-19

P

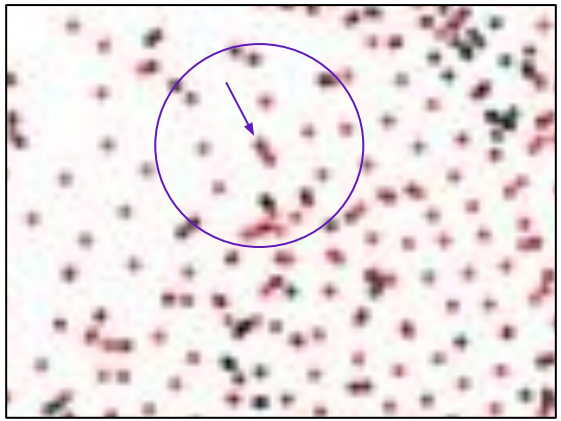
Update

C

~~Update~~

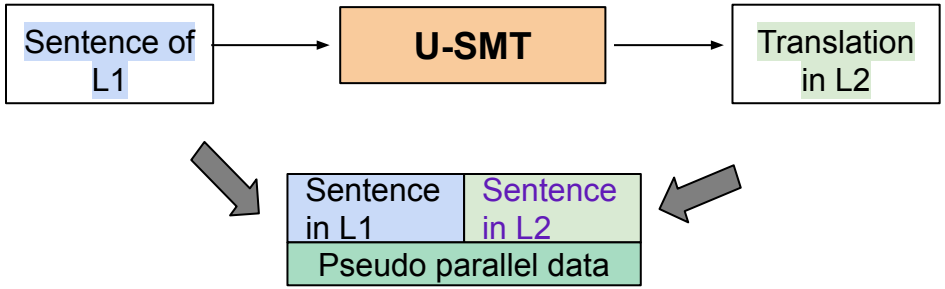
# Phrase table induction in an unsupervised way

- Get cross-lingual n-gram embedding.
- Calculate Phrase-translation probabilities.
  - Limit the translation candidates for each source phrase to its 100 nearest neighbors in the target language.
  - Apply the softmax function over the cosine similarities of their respective embeddings.



# Unsupervised Tuning

- Tuning with synthetic data.
  - Generate a synthetic parallel corpus.
  - Apply MERT tuning over it iteratively repeating the process in both directions.




- Unsupervised optimization objective:
  - Cyclic loss: The translation of translation of a sentence should be close to the original text.
  - LM loss: We want a fluent sentence in the target language.

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

# Iterative refinement

- Generate a synthetic parallel corpus by translating the monolingual corpus with the initial system L1→L2, and train and tune SMT system L2→L1.
  - To accelerate the experiments, use a random subset of 2 million sentences from each monolingual corpus for training.
  - Reuse the original language model, which is trained in the full corpus.
- The process can be repeated iteratively until some convergence criterion is met.

# Adding subword information

- We want to favor phrase translation candidates that are similar at the character level.
- Additional weights are added to initial phrase-table.
  - Unlike lexical weightings it use a character-level similarity function instead of word translation probabilities.

$$\text{score}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j \text{sim}(\bar{f}_i, \bar{e}_j)\right)$$

# Results

| | WMT-14 | | | | WMT-16 | |
|---|---|---|---|---|---|---|
| | **FR-EN** | **EN-FR** | **DE-EN** | **EN-DE** | **DE-EN** | **EN-DE** |
| Unsupervised SMT | 21.16 | 20.13 | 13.86 | 10.59 | 18.01 | 13.22 |
| + unsupervised tuning | 22.17 | 22.22 | 14.73 | 10.64 | 18.21 | 13.12 |
| + iterative refinement (it1) | 24.81 | 26.53 | 16.01 | 13.45 | 20.76 | 16.94 |
| + iterative refinement (it2) | **26.13** | **26.57** | 17.30 | 13.95 | 22.80 | 18.18 |
| + iterative refinement (it3) | 25.87 | 26.22 | **17.43** | **14.08** | **23.05** | **18.23** |

Artetxe, M., Labaka, G. and Agirre, E., 2018. Unsupervised Statistical Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3632-3642).

# UMT Approaches

1. Unsupervised NMT
2. GAN for UNMT
3. Unsupervised SMT
4. **Hybrid UMT**

# Introduction

- We can combine UNMT and USMT in two ways.
  - USMT followed by UNMT.
  - UNMT followed by USMT.

List of papers


1. Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.A., 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049).

2. Artetxe, M., Labaka, G. and Agirre, E., 2019, July. An Effective Approach to Unsupervised Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 194-203).

# USMT followed by UNMT Vs. UNMT followed by USMT

- USMT followed by UNMT:
  - Generate pseudo parallel data with USMT.
  - Initialise UNMT system with the pseudo parallel data.
- UNMT followed by USMT:
  - Generate pseudo parallel data with UNMT.
  - Initialise USMT system with the pseudo parallel data.

USMT followed by UNMT wins.

| WMT 14/16 | En→Fr | Fr→En | En→De | De→En |
|---|---|---|---|---|
| NMT + PBSMT | 27.1 | 26.3 | 17.5 | 22.1 |
| PBSMT + NMT | **27.6** | **27.7** | **20.2** | **25.2** |

Lample, G., Ott, M., Conneau, A., Denoyer, L. and Ranzato, M.A., 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5039-5049).

# Pre-training approaches for Unsupervised NMT

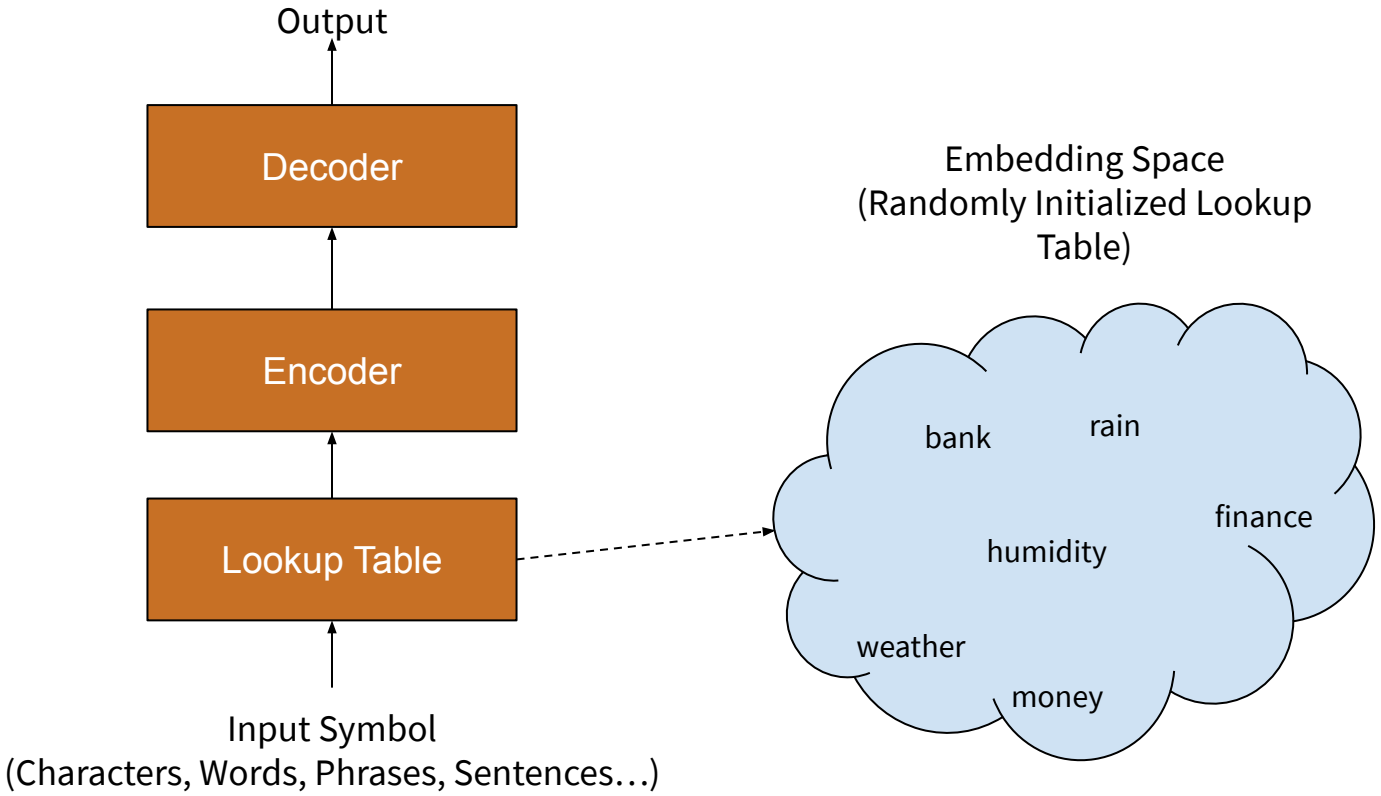
# XLM, CMLM, MASS, BART, mBART

# XLM

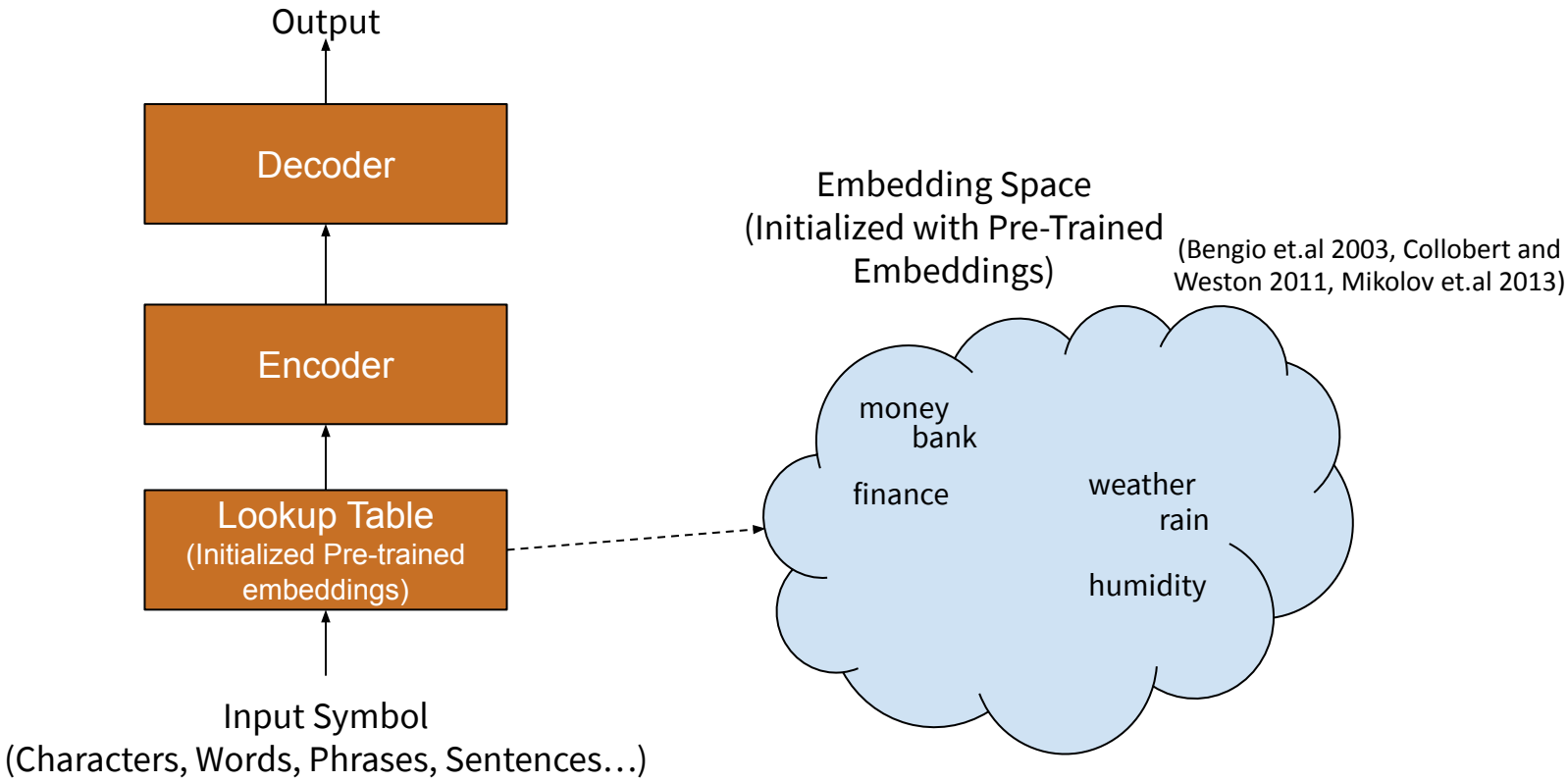
Cross-lingual Language Modelling Pre-Training

Cross-lingual Language Model Pretraining, Advances in Neural Information Processing Systems. 2019.

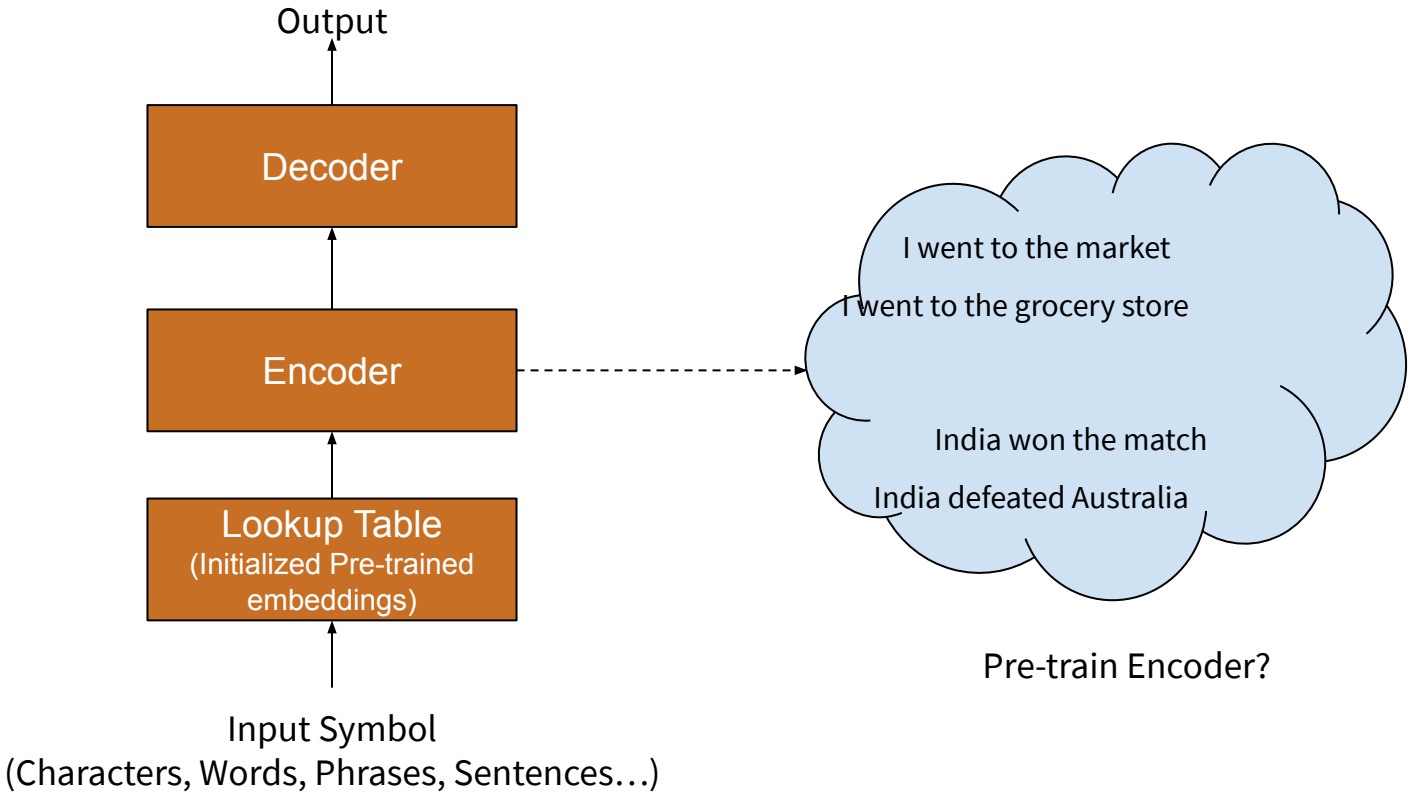
# Typical Deep Learning Module



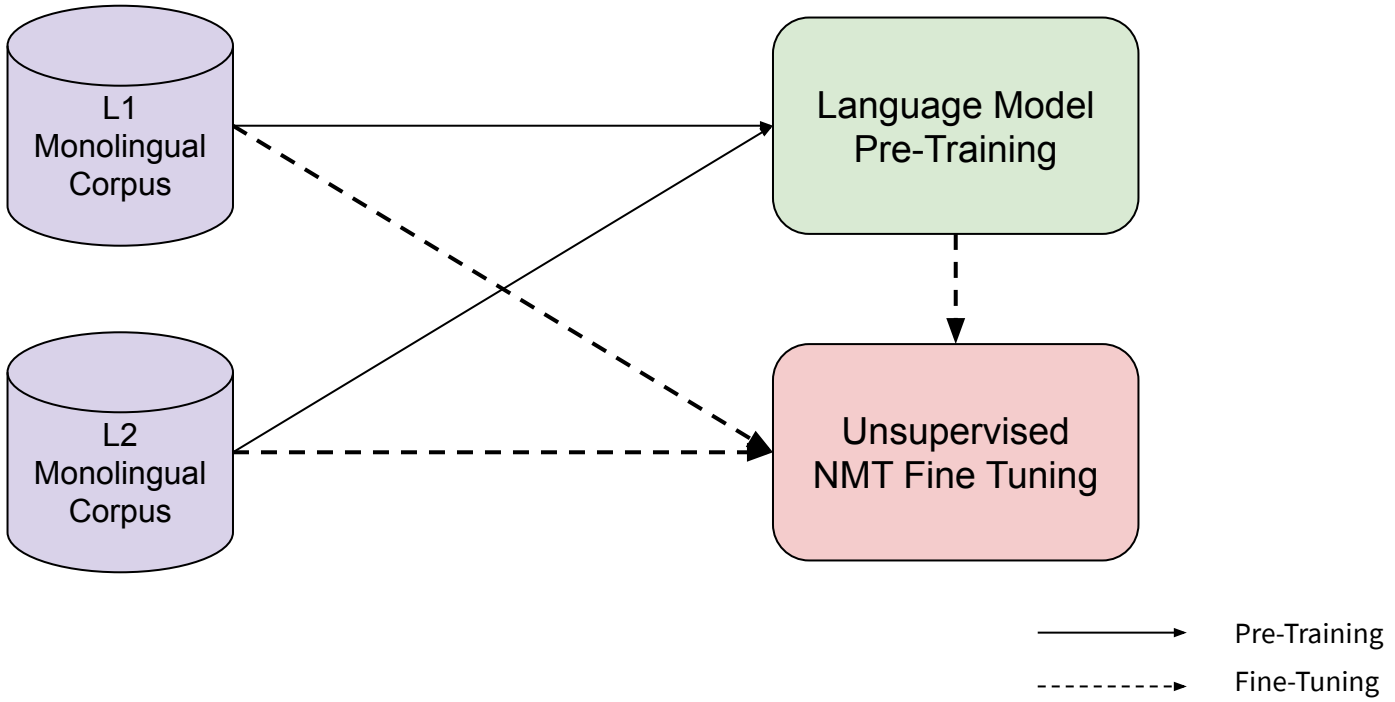
# Typical Deep Learning Module



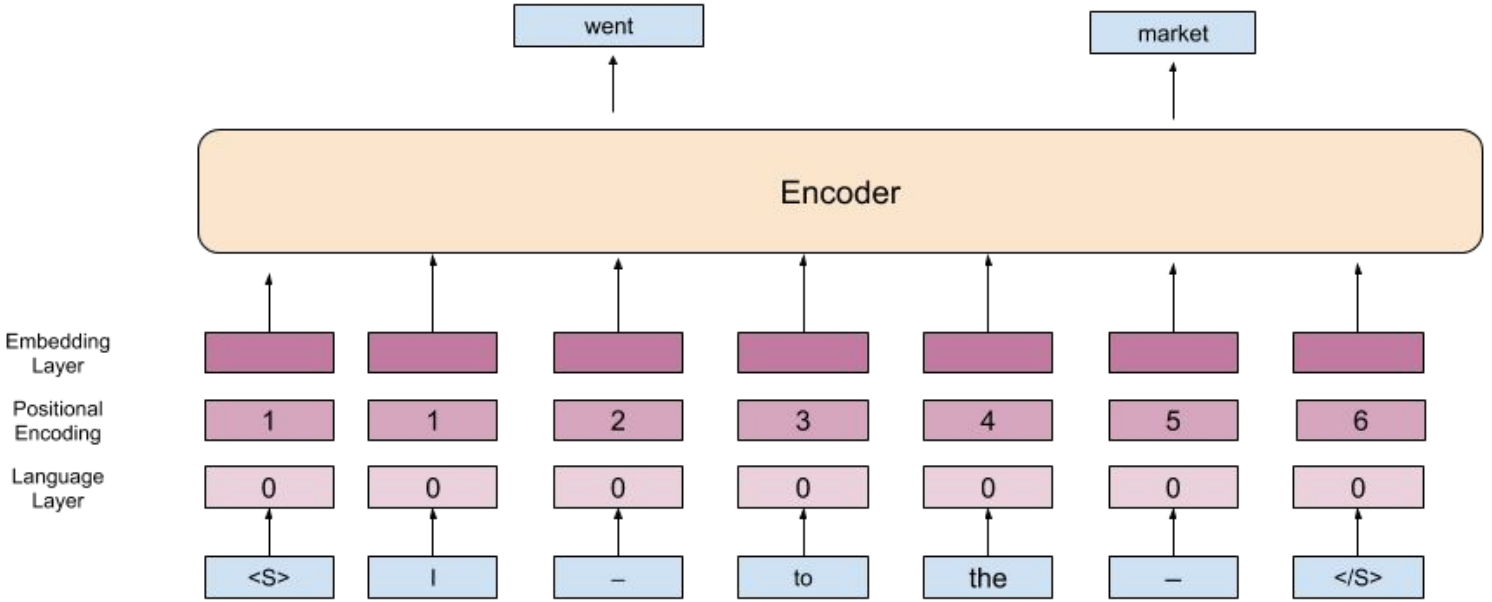
# Typical Deep Learning Module



# General Framework



# XLM Pre-Training



# XLM Fine Tuning

- Perform fine-tuning using
  - Iterative back-translation
  - Denoising auto-encoding
- Alternate between the two objective
- Denoising auto-encoding helps in better training of the decoder

# XLM: Results

| | | en-fr | fr-en | en-de | de-en | en-ro | ro-en |
|---|---|---|---|---|---|---|---|
| *Previous state-of-the-art - Lample et al. (2018b)* | | | | | | | |
| NMT | | 25.1 | 24.2 | 17.2 | 21.0 | 21.2 | 19.4 |
| PBSMT | | 28.1 | 27.2 | 17.8 | 22.7 | 21.3 | 23.0 |
| PBSMT + NMT | | 27.6 | 27.7 | 20.2 | 25.2 | 25.1 | 23.9 |
| *Our results for different encoder and decoder initializations* | | | | | | | |
| EMB | EMB | 29.4 | 29.4 | 21.3 | 27.3 | 27.5 | 26.6 |
| - | - | 13.0 | 15.8 | 6.7 | 15.3 | 18.9 | 18.3 |
| - | CLM | 25.3 | 26.4 | 19.2 | 26.0 | 25.7 | 24.6 |
| - | MLM | 29.2 | 29.1 | 21.6 | 28.6 | 28.2 | 27.3 |
| CLM | - | 28.7 | 28.2 | 24.4 | 30.3 | 29.2 | 28.0 |
| CLM | CLM | 30.4 | 30.0 | 22.7 | 30.5 | 29.0 | 27.8 |
| CLM | MLM | 32.3 | 31.6 | 24.3 | 32.5 | 31.6 | 29.8 |
| MLM | - | 31.6 | 32.1 | **27.0** | 33.2 | 31.8 | 30.5 |
| MLM | CLM | **33.4** | 32.3 | 24.9 | 32.9 | 31.7 | 30.4 |
| MLM | MLM | **33.4** | **33.3** | 26.4 | **34.3** | **33.3** | **31.8** |

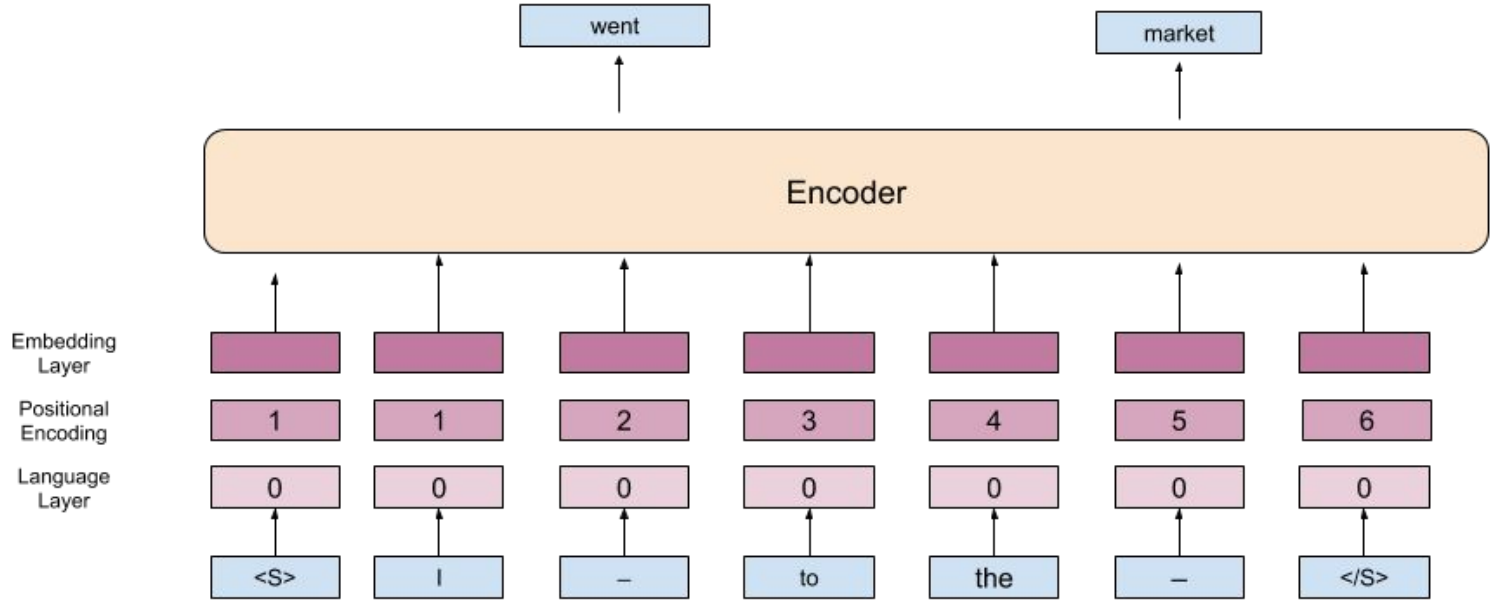
- MLM objective results in better BLEU score compared to Causal Language Modeling (CLM) objective

# CMLM

Cross-lingual Masked Language Modelling

Explicit Cross-lingual Pre-training for Unsupervised Machine Translation, EMNLP-IJCNLP 2019

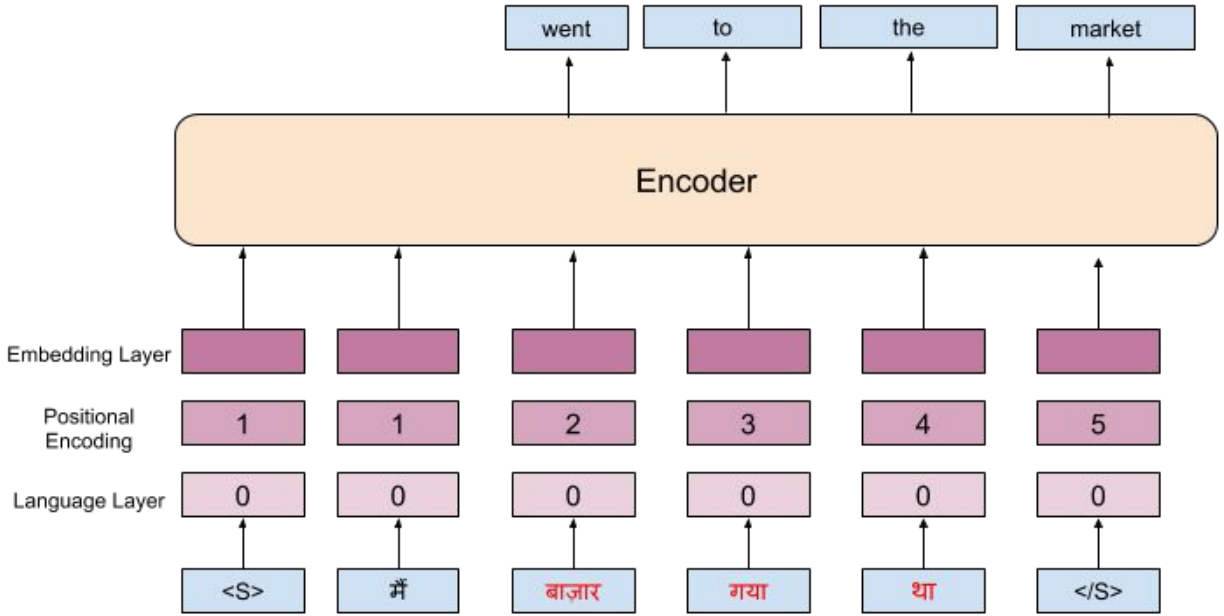
# MLM (Devlin et.al 2018)



# Limitations

- MLM is trained to predict the missing word in the sentence
- Also, joint training on the combined corpus is not a strong signal to learn good multilingual representations
- Provide explicit cross-lingual signals to the model while pre-training

# Cross-lingual Masked Language Modelling



# Cross-lingual Masked Language Modelling

- Obtain n-gram phrase translations as discussed earlier
- MLM tries to predict the masked words/tokens
- Modify MLM objective to predict the translation of phrases
- **Mismatch between source and target phrase length**

# Cross-lingual Masked Language Modelling

**Challenges**

- The source and target phrases are of unequal length
- For BERT or XLM, the decoder is a linear classifier.
- Introduce IBM model-2 into the objective

$$P(y_1^m \mid x_1^l) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i)$$

$\epsilon$ = probability that the translation of $x_1^l$ consists of m tokens

$a(i, |j, l, m)$ = probability that $i^{th}$ source token is aligned to $j^{th}$ target token

# Cross-lingual Masked Language Modelling

**Modeling**

- Introduce IBM model-2 into the objective

$$P(y_1^m \mid x_1^l) = \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i)$$

$\epsilon$ = probability that the translation of $x_1^l$ consists of m tokens
$a(i, |j, l, m)$ = probability that $i^{th}$ source token is aligned to $j^{th}$ target token

- The loss function becomes

$$L_{cmlm} = -\log(\epsilon) - \sum_{j=1}^{m} \log\left(\sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i)\right)$$

# Cross-lingual Masked Language Modelling

**Modeling**

- The loss function becomes

$$L_{cmlm} = -\log(\epsilon) - \sum_{j=1}^{m} \log\left(\sum_{i=0}^{l} a(i, |j, l, m)\, P(y_j \mid x_i)\right)$$

- The gradient becomes:

$$\nabla L = \sum_{j=1}^{m} \frac{a(i \mid j, l, m)\, P(y_j \mid x_i)}{\sum_{i=0}^{l} a(i \mid j, l, m)\, P(y_j \mid x_i)} \nabla \log P(y_j \mid x_i)$$

# Cross-lingual Masked Language Modelling

**Modeling**

- The gradient becomes:

$$\nabla L = \sum_{j=1}^{m} \frac{a(i \mid j, l, m)\, P(y_j \mid x_i)}{\sum_{i=0}^{l} a(i \mid j, l, m)\, P(y_j \mid x_i)} \nabla \log P(y_j \mid x_i)$$

  - a(i, |j, l, m) are approximated using cross-lingual BPE embedding
  - $P(y_j \mid x_i)$ is calculated by passing $x_i$ contextual embedding representation through a linear layer followed by soft-max

# Cross-lingual Masked Language Modelling

**Algorithm**

- Alternate between CMLM and MLM objective
- In MLM objective,
  - 50% of the time randomly choose some source ngrams  and replace it with the corresponding translation candidate (pseudo code-switching)
- In CMLM objective,
  - Randomly select 15% of the BPE ngram tokens and replace them by [MASK] 70% of the time
  - Trained to predict the translation candidate in the other language

# Cross-lingual Masked Language Modelling

**Results**

| Method | fr2en | en2fr | de2en | en2de | ro2en | en2ro |
|---|---|---|---|---|---|---|
| (Artetxe et al., 2017) | 15.6 | 15.1 | - | - | - | - |
| (Lample et al., 2017) | 14.3 | 15.1 | 13.3 | 9.6 | - | - |
| (Artetxe et al., 2018b) | 25.9 | 26.2 | 23.1 | 18.2 | - | - |
| (Lample et al., 2018) | 27.7 | 28.1 | 25.2 | 20.2 | 23.9 | 25.1 |
| (Ren et al., 2019) | 28.9 | 29.5 | 26.3 | 21.7 | - | - |
| (Lample and Conneau, 2019) | 33.3 | 33.4 | 34.3 | 26.4 | 31.8 | 33.3 |
| Iter 1 (CMLM) | 34.8 | 34.9 | 35.5 | **27.9** | 33.6 | 34.7 |
| Iter 2 (CMLM) | **34.9** | **35.4** | **35.6** | 27.7 | **34.1** | **34.9** |

# CMLM

Cross-lingual Masked Language Modelling

Ablation Study

# CMLM: Ablation Study

- Role of n-gram masking
- Influence of translation prediction

| | fr2en | en2fr | de2en | en2de |
|---|---|---|---|---|
| CMLM + MLM | **34.8** | **34.9** | **35.5** | **27.9** |
| CMLM | 34.1 | 34.3 | 35.1 | 27.2 |
| - translation prediction | 33.7 | 33.9 | 34.8 | 26.6 |
| - - n-gram mask | 33.3 | 33.4 | 34.3 | 26.4 |

**CMLM + MLM** means we use $L_{pre}$ as the pre-training loss;
**CMLM** means we only use $L_{cmlm}$ as the pre-training loss;
**-- translation prediction** predict the masked n-grams rather than their translation candidates;
**- - n-gram mask** randomly mask BPE tokens rather than n-grams based on **-- translation prediction** during pre-training, which degrades our method to XLM.

# MASS

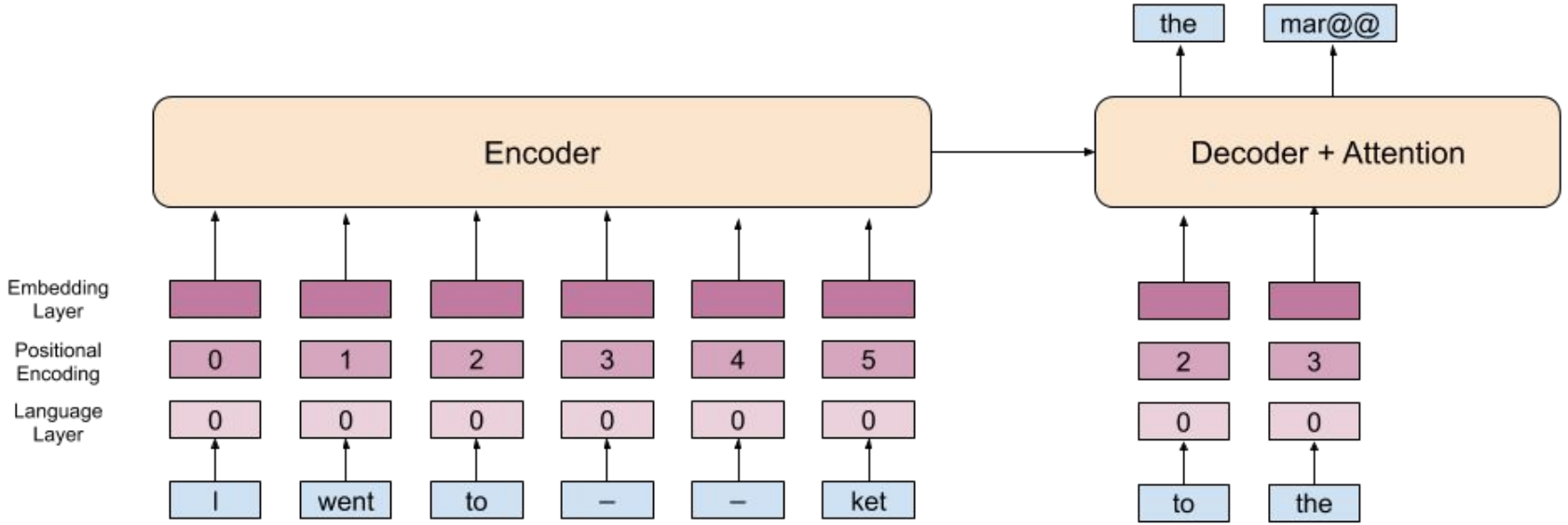
Masked Sequence to Sequence pretraining

MASS: Masked Sequence to Sequence Pre-training for Language Generation, ICML, Song et.al 2019

# MASS (Song et.al 2019)

- XLM objective predicts the masked word in the sentence
- However, for U-NMT we need to generate a sequence
- This disconnect between pre-training and fine-tuning objective could limit the potential of unsupervised pre-training
- MASS extends XLM objective to include text segments
- Given a sentence, randomly mask k% of the text segment
- The decoder has to generate the masked text segment now

# MASS Pre-Training



# MASS Fine-Tuning

- Perform fine-tuning using iterative back-translation
- Unlike XLM which had
  - iterative back-translation
  - Denoising auto-encoding

# MASS (Song et.al 2019)

| Method | Setting | en - fr | fr - en | en - de | de - en | en - ro | ro - en |
|---|---|---|---|---|---|---|---|
| Artetxe et al. (2017) | 2-layer RNN | 15.13 | 15.56 | 6.89 | 10.16 | - | - |
| Lample et al. (2017) | 3-layer RNN | 15.05 | 14.31 | 9.75 | 13.33 | - | - |
| Yang et al. (2018) | 4-layer Transformer | 16.97 | 15.58 | 10.86 | 14.62 | - | - |
| Lample et al. (2018) | 4-layer Transformer | 25.14 | 24.18 | 17.16 | 21.00 | 21.18 | 19.44 |
| XLM (Lample & Conneau, 2019) | 6-layer Transformer | 33.40 | 33.30 | 27.00 | 34.30 | 33.30 | 31.80 |
| MASS | 6-layer Transformer | **37.50** | **34.90** | **28.30** | **35.20** | **35.20** | **33.10** |

*Table 2.* The BLEU score comparisons between MASS and the previous works on unsupervised NMT. Results on en-fr and fr-en pairs are reported on *newstest2014* and the others are on *newstest2016*. Since XLM uses different combinations of MLM and CLM in the encoder and decoder, we report the highest BLEU score for XLM on each language pair.

# MASS

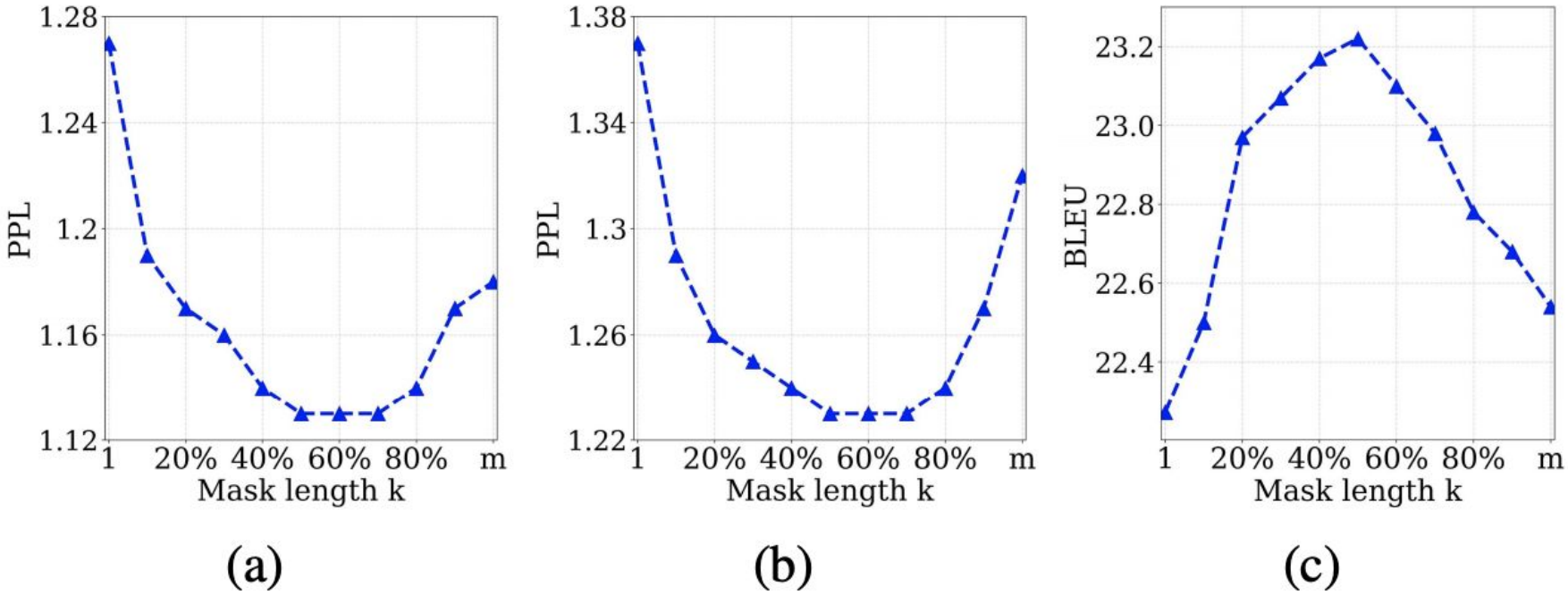
Masked Sequence to Sequence pretraining

Role of hyper-parameters

# MASS Hyper-parameters

- Percentage of ngram tokens in a sentence to be masked (**masking length**)
  - Consider the input sentence, **X** = ***I went to the market yesterday night***
  - Let ***to the market yesterday*** be the text segment selected for masking
  - Default value is 50% of the input sentence
  - However, not all tokens ***to the market yesterday*** are masked
- Given a text fragment $x_i, \ldots, x_j$ of length m selected for masking (Word selection)
  - k% of the tokens are selected for masking (**mask probability**)
  - l% of the tokens are replaced by random tokens (**replace probability**)
  - ( 100 - (k + l) ) of the tokens are retained (**keep probability**)
  - Default values are k = 80%, l = 10%

# MASS (Song et.al 2019): Role of Masking Length




The performances of MASS with different masked lengths k, in both pre-training and fine-tuning stages, which include: the PPL of the pre-trained model on English (Figure a) and French (Figure b) sentences from WMT newstest2013 on English-French translation; the BLEU score of unsupervised English-French translation on WMT newstest2013 (Figure c)

# MASS Hyper-parameters

मै तो आपके घर से चाय का पत्ती मांगने आयी  हूँ

mai to Apake ghara se chAya kA pattI mAMgane AyI  hU.N

Select randomly 50% of the consecutive tokens for masking

मै तो आपके _ _ _ _ _ मांगने आयी  हूँ

80% of the selected tokens are **masked**, 10% **randomly replaced**

मै तो आपके _ से _ पीना _ मांगने आयी  हूँ

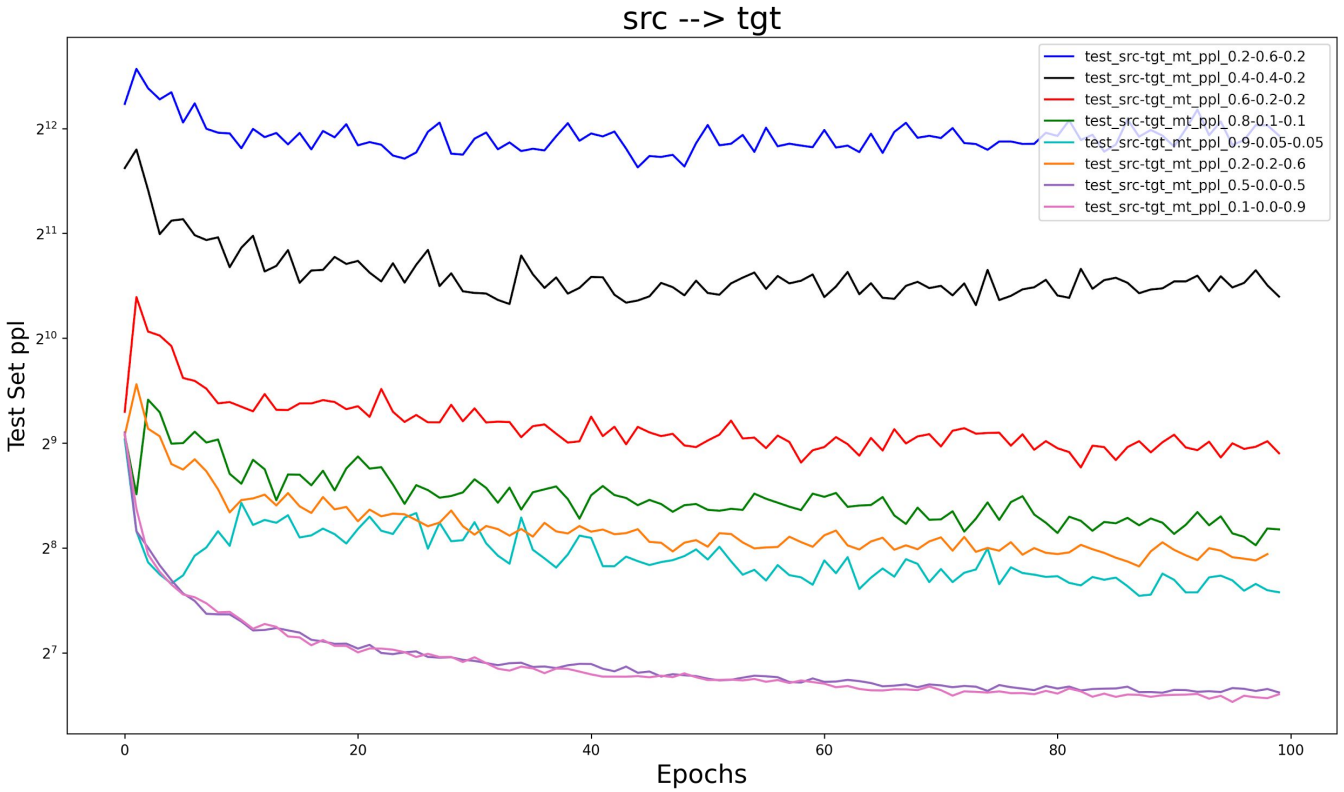
घर से चाय का पत्ती

Output to be generated

# MASS: Word Selection Hyper-parameters

| Configuration | %age Masked | %age Retained | %age Randomly replaced |
|---|---|---|---|
| 1 | 20 | 60 | 20 |
| 2 | 40 | 40 | 20 |
| 3 | 60 | 20 | 20 |
| 4 | 80 | 10 | 10 |
| 5 | 90 | 5 | 5 |
| 6 | 20 | 20 | 60 |
| 7 | 50 | - | 50 |
| 8 | 10 | - | 90 |

# MASS (Song et.al 2019): Word Selection Hyper-parameters



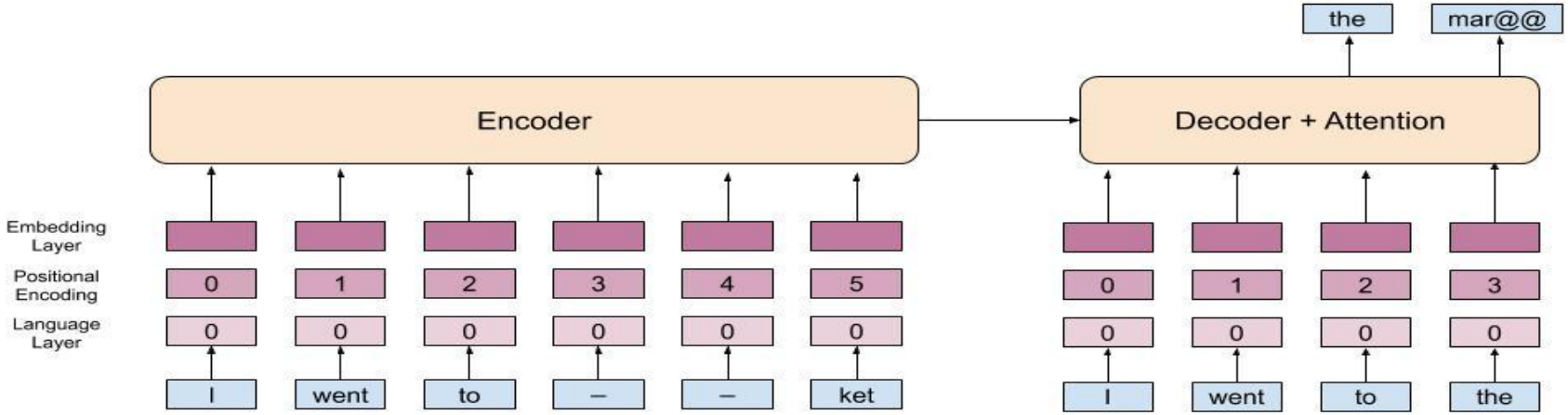
# MASS: Word Selection Hyper-parameters

| Configuration | %age Masked | %age Retained | %age Randomly replaced | Comments |
|---|---|---|---|---|
| 1 | 20 | 60 | 20 | Auto-encoder |
| 2 | 40 | 40 | 20 | Auto-encoder |
| 3 | 60 | 20 | 20 | Auto-encoder |
| 4 | 80 | 10 | 10 | Recommended |
| 5 | 90 | 5 | 5 | Recommended |
| 6 | 20 | 20 | 60 | Unable to generate translations. But perplexity is low (Better for other tasks?) |
| 7 | 50 | - | 50 | |
| 8 | 10 | - | 90 | |

# MASS (Song et.al 2019): Role of Masking Tokens

- Consider the input sentence, **X** = ***I went to the market yesterday night***
- Let ***to the market yesterday*** be the text segment selected for masking
- The input to the encoder is ***I went _ _ _ _ night***
- The input to the decoder (previous token) is ***went to the market***
  - Why mask consecutive tokens and not discrete tokens? (***Discrete***)
  - Why not feed all the input tokens to the decoder (similar to previous target word in NMT)? (***feed***)

# Feeding Input Tokens



# MASS (Song et.al 2019): Role of Masking Tokens

| Method | BLEU | Method | BLEU | Method | BLEU |
|---|---|---|---|---|---|
| *Discrete* | 36.9 | *Feed* | 35.3 | MASS | 37.5 |

The comparison between MASS and the ablation methods in terms of BLEU score on the unsupervised en-fr translation

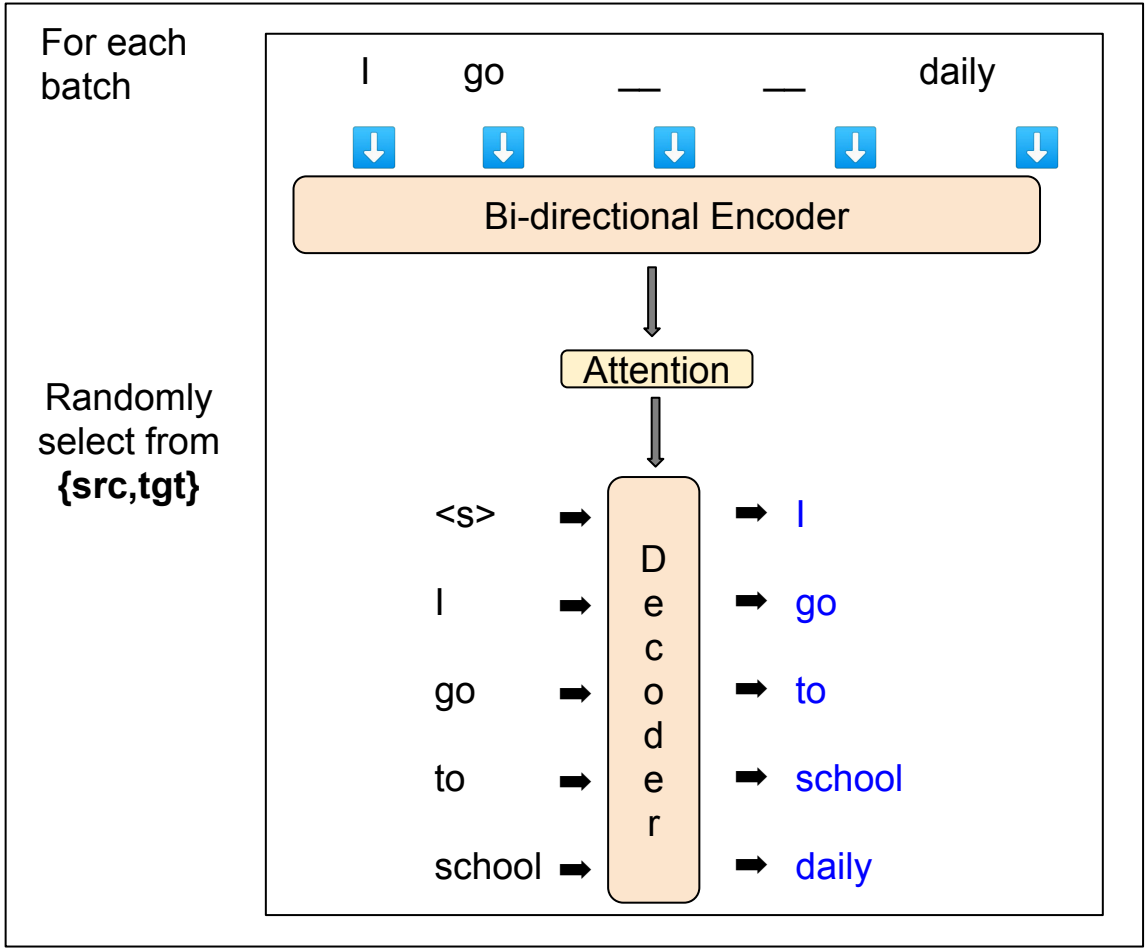
# BART and mBART

BART: Denoising Sequence to Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, ACL 2020, (Lewis et al 2020)

Multilingual denoising pre-training for Neural Machine Translation, 2020, (Liu et al 2020)

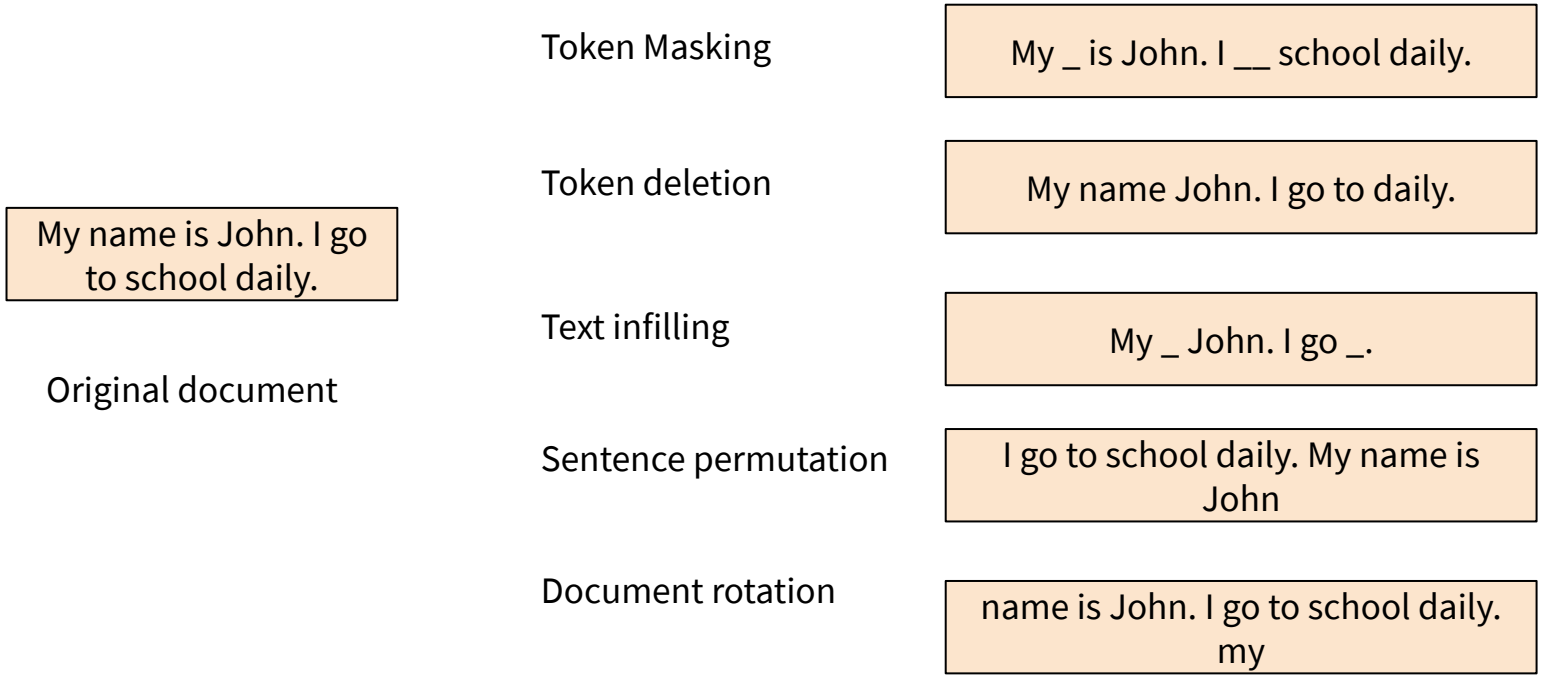
# BART Pretraining

- Trained by
  - Corrupting text with an arbitrary noising function
  - Learning a model to reconstruct the original text.
- Denoising full text
- Multi-sentence level




Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL 2020)*

# BART pretraining (possible noising steps) (Lewis et al. 2020)



# BART noising steps (Lewis et al. 2020)

- Experimented with different noise functions for various tasks
  - Text infilling + Sentence permutation performed the best
    - Remove spans of text and replace with mask tokens
    - Mask 30% of the words in each instance by randomly sampling a span length
    - Permute the order of sentences

# mBART (Liu et al 2020)

- A sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in **many languages** using the BART objective
- Unsupervised NMT
  - BART pretraining using monolingual corpora of multiple languages + Iterative Back-Translation

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv2020.

# mBART (Liu et al 2020)

- Pre-training using BART objective on multiple languages

| Model | Similar Pairs | | | | Dissimilar Pairs | | | |
|---|---|---|---|---|---|---|---|---|
| | En-De | | En-Ro | | En-Ne | | En-Si | |
| | ← | → | ← | → | ← | → | ← | → |
| **Random** | 21.0 | 17.2 | 19.4 | 21.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| **XLM (2019)** | 34.3 | 26.4 | 31.8 | 33.3 | 0.5 | 0.1 | 0.1 | 0.1 |
| **MASS (2019)** | **35.2** | 28.3 | **33.1** | **35.2** | - | - | - | - |
| **mBART** | 34.0 | **29.8** | 30.5 | 35.0 | **10.0** | **4.4** | **8.2** | **3.9** |

- En-De and En-ro are only trained using specified source and target languages
- En-Ne and En-Si, the pretraining is performed using mBART on 25 languages.
- mBART also generalizes well for the languages not seen in pretraining.

Results: mBART (only on source and target language) pretraining for unsupervised NMT

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv preprint arXiv:2001.08210, 2020.

# When Unsupervised NMT does not work?

Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? arXiv preprint

# Factors impacting the performance of Unsupervised NMT

- **Domain similarity**
  - Sensitive to domain mismatch
- **Dissimilar language pairs**
  - The similarity between language pairs helps the model in training good shared encoder
- **Initial model to start pretraining**
  - Good initializations leads to good performance in the finetuning phase
- **Unbalanced data size**
  - Not useful to use oversized data on one side
- **Quality of cross-lingual embeddings**
  - Initialization is done using cross-lingual embeddings

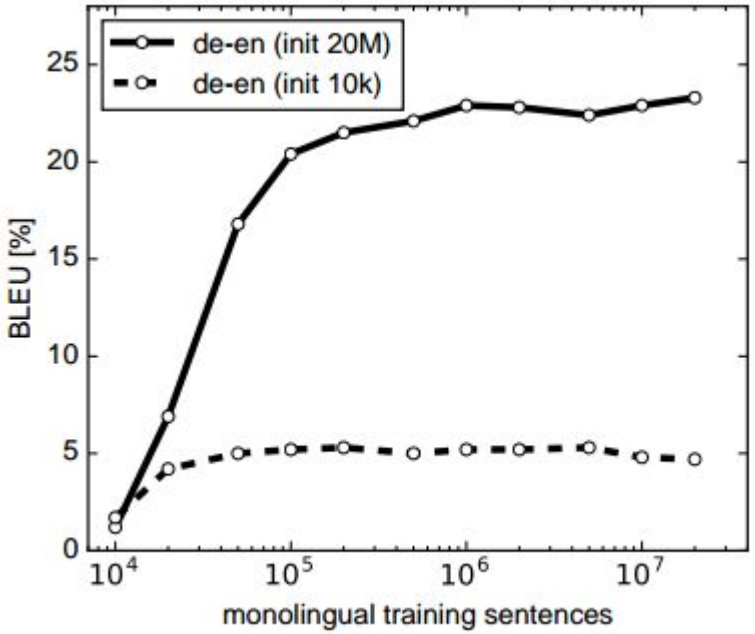
# Domain similarity

| Domain (en) | Domain (de/ru) | BLEU [%] de-en | en-de | ru-en | en-ru |
|---|---|---|---|---|---|
| Newswire | Newswire | 23.3 | 19.9 | 11.9 | 9.3 |
| | Politics | 11.5 | 12.2 | 2.3 | 2.5 |
| | Random | 18.4 | 16.4 | 6.9 | 6.1 |

- Different distributions of the topics

Image source: Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

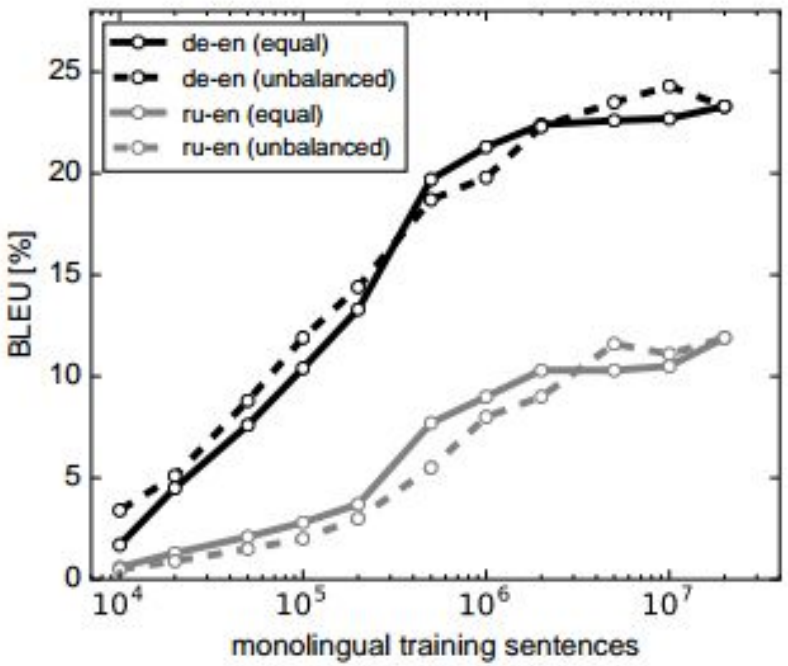
# Initialization




- Good initializations leads to good performance in the fine-tuning phase
- Final model correlates well with the initialization quality

Image source: Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

# Unbalanced data size




Target side training data: 20M sentences

Solid line: target data has the same number of source and target sentences

- Not useful to use oversized data on one side

Image source: Graça, Yunsu Kim Miguel, and Hermann Ney. "When and Why is Unsupervised Neural Machine Translation Useless?." In 22nd Annual Conference of the European Association for Machine Translation, p. 35.

# Quality of Cross-lingual Embeddings

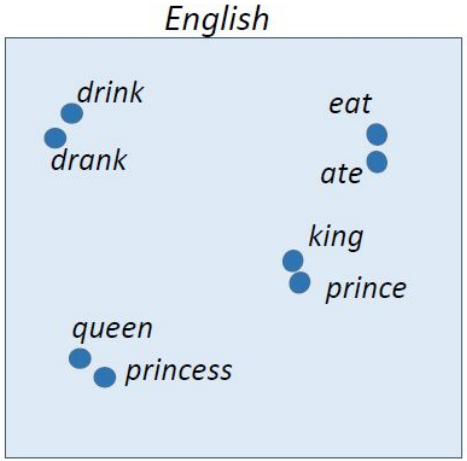
# Cross-lingual Word Embeddings: Quality?

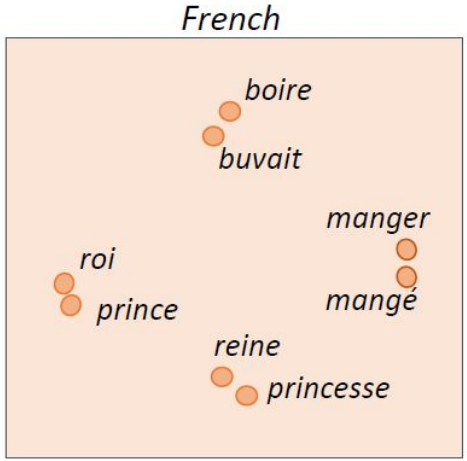
## Unsupervised NMT [Lample et al 2018]

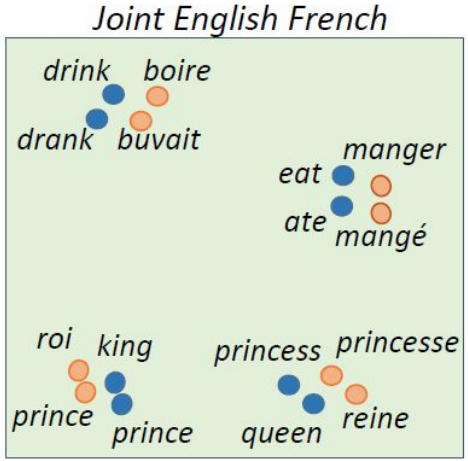
Pre-processing

1. Obtain cross-lingual embeddings either in an unsupervised manner or supervised manner
2. The pre-trained cross-lingual embeddings are not updated during training
3. Success of the approach relies on the quality of cross-lingual embeddings in addition to other factors like *language relatedness, etc*

# Cross-lingual Representations



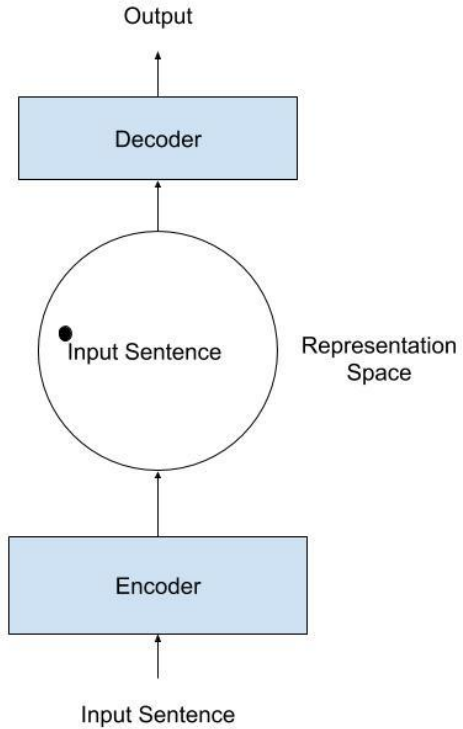




Monolingual Word Representations
*(capture syntactic and semantic similarities between words)*

Multilingual Word Representations
*(capture syntactic and semantic similarities between words both within and across languages)*

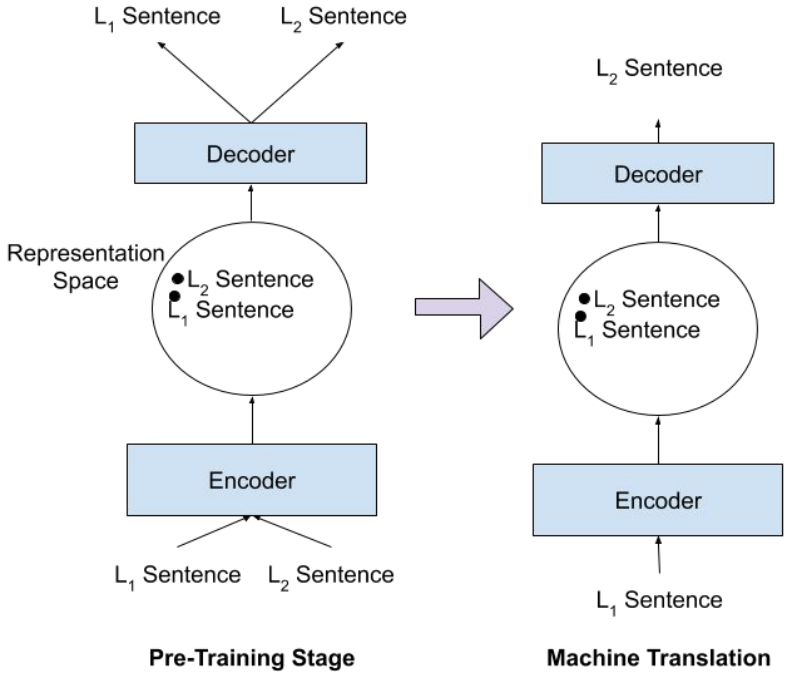
*(Source: Khapra and Chandar, 2016)*

# Why is the Quality questioned?




Encode-Decode paradigm used for MT

# Good Quality Cross-lingual Embeddings?




The ability of the encoder to learn better multilingual representations lies on the quality of cross-lingual embeddings

Encode-Decode paradigm used for MT

# Quantitative Quality

| Source - Target | GeoMM |
|---|---|
| En - Es | 81.4 |
| Es - En | 85.5 |
| En - Fr | 82.1 |
| Fr - En | 84.1 |
| En - De | 74.7 |
| De - En | 76.7 |
| En - Hi | 41.5 |
| Hi - En | 54.8 |
| En - Ta | 31.9 |
| Ta - En | 38.7 |
| En - Bn | 36.7 |
| Bn - En | 42.7 |

Very low Precision@1 for Indic languages compared to the European language counterpart

Precision@1 for BLI task using GeoMM on MUSE dataset (*Jawapuria et.al 2019, Kakwani et.al 2020*)

# Unsupervised NMT [Lample et al 2018]

Simple word-by-word translation using cross-lingual embeddings

| | Multi30k-Task1 | | | | WMT | | | |
|---|---|---|---|---|---|---|---|---|
| | en-fr | fr-en | de-en | en-de | en-fr | fr-en | de-en | en-de |
| Supervised | 56.83 | 50.77 | 38.38 | 35.16 | 27.97 | 26.13 | 25.61 | 21.33 |
| word-by-word | 8.54 | 16.77 | 15.72 | 5.39 | 6.28 | 10.09 | 10.77 | 7.06 |
| word reordering | - | - | - | - | 6.68 | 11.69 | 10.84 | 6.70 |

# Unsupervised NMT [Lample et al 2018]

Simple word-by-word translation using cross-lingual embeddings

| Language Pair | BLEU Score |
|---|---|
| En → Fr | 6.28 |
| Fr → En | 10.09 |
| En → De | 7.06 |
| De → En | 10.77 |
| **En → Hi** | **1.2** |
| **Hi → En** | **2.1** |

**Credit:** Tamali for the English-Hindi numbers

# Cross-lingual Embedding Quality

1. Poor Cross-lingual Embeddings leads to diminished returns from U-NMT methods

# Future Directions

1. Learn better cross-lingual embeddings between Indic languages and Indic to European languages
2. Majority of the NLP approaches operate at sub-word level
3. How to obtain cross-lingual embeddings at the sub-word level?

# Unsupervised NMT for Indic languages

Initial Findings

# Why Indic Languages?

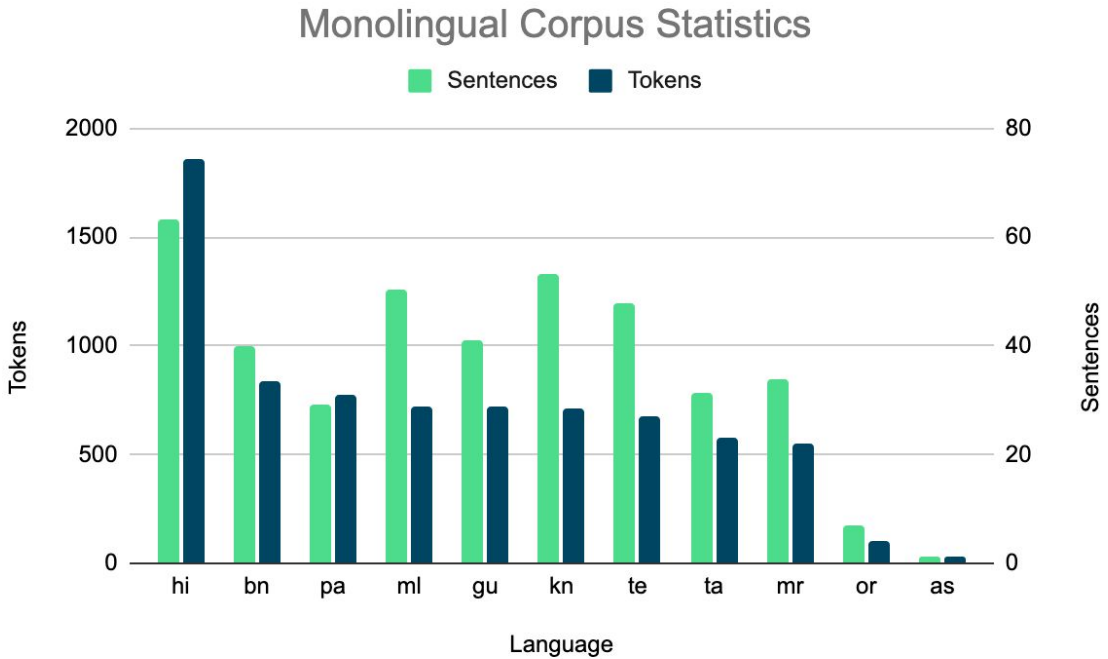
- A test-bed for research on multilinguality
- Spectrum of language similarity

| | Bn | Gu | Hi | Mr | Pa | Ml | Ta | Te |
|---|---|---|---|---|---|---|---|---|
| **Bn** | - | 19.51 | 29.45 | 11.39 | 2.45 | 1.05 | 0.34 | 0.78 |
| **Gu** | 13.9 | - | 51.75 | 20.14 | 4.46 | 1.06 | 0.3 | 1.22 |
| **Hi** | 12.76 | 31.47 | - | 15.22 | 4.43 | 0.78 | 0.21 | 0.95 |
| **Mr** | 11.81 | 29.31 | 36.42 | - | 3.4 | 0.62 | 0.27 | 0.92 |
| **Pa** | 4.26 | 10.88 | 17.79 | 5.71 | - | 0.22 | 0.16 | 0.4 |
| **Ml** | 1.19 | 1.7 | 2.04 | 0.67 | 0.14 | - | 0.72 | 2.48 |
| **Ta** | 0.43 | 0.54 | 0.62 | 0.33 | 0.11 | 0.8 | - | 0.25 |
| **Te** | 0.95 | 2.1 | 2.67 | 1.08 | 0.28 | 2.68 | 0.24 | - |

Percentage of words in the source language (row) which also appear in the target language (column) (transliterated to a common script) and having at least one common synset obtained from Indo-Wordnet (Bhattacharyya et.al 2010)

# Why Indic Languages?

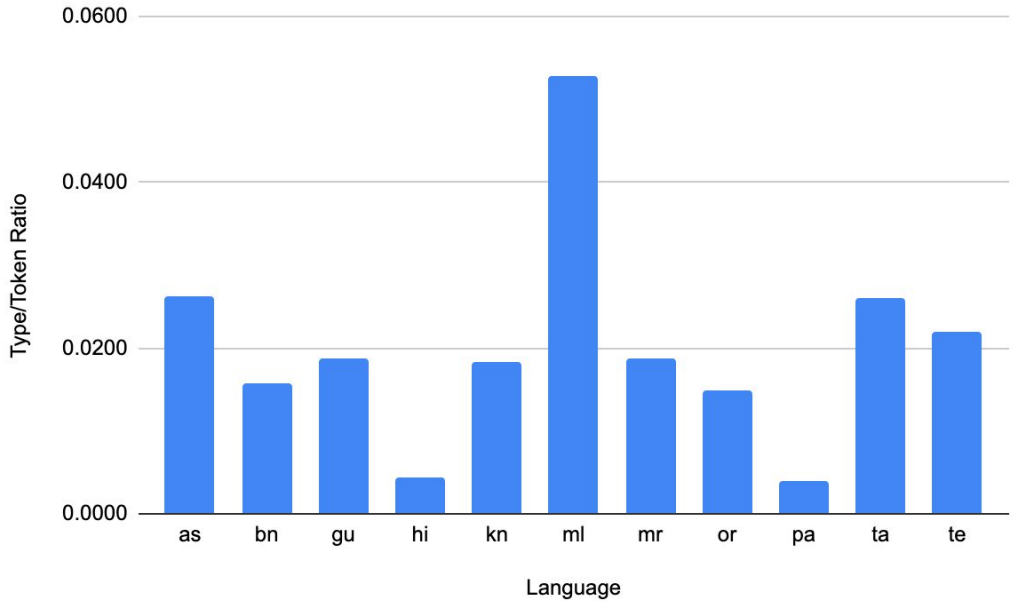
- Low-resourceness




Monolingual Corpus Statistics (in Millions) (Kunchukuttan et.al 2020)

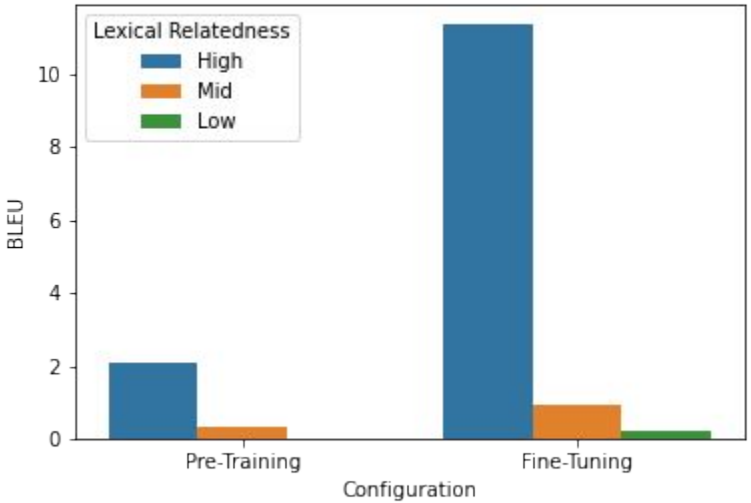
# Why Indic Languages?

- Spectrum of morphological complexity

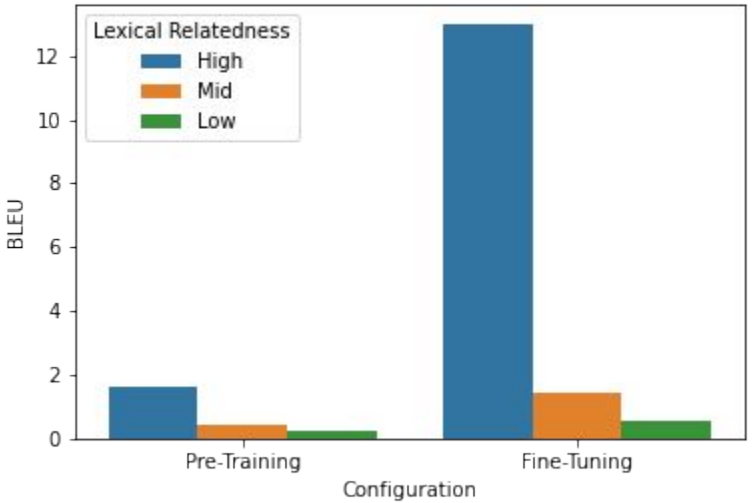



Type-Token Ratio calculated on AI4Bharat Corpus (Kunchukuttan et.al 2020)

# U-NMT for Indic Languages: Results




Source → Target




Target → Source

# Conclusions

1. Existing U-NMT models fail for Indic languages
2. For closely-related languages, we observe decent BLEU scores
3. Morphological richness adds more complexity to the model
4. Need **more research** focusing on **Indic languages**

# Conclusions

# Conclusion

- Paradigms of the MT task.
- Foundational concepts for the U-NMT paradigm.
- U-NMT approaches.
- Recent language modeling approaches.
- Results for Indian language pairs (related and unrelated languages).
- Need for further research in the area of U-NMT.

# Future of U-NMT

1. U-NMT approaches have shown promising results for closely-related languages
2. U-NMT performs poor for distant languages
3. Better cross-lingual embeddings for distant languages.
4. Better cross-lingual language model pretraining for resource-scarce languages, disimilar languages, and dissimilar domans

# Resources

- Resources can be found here

  www.cfilt.iitb.ac.in

- The tutorial slides will be uploaded here

  https://github.com/murthyrudra/unmt_tutorial_icon2020

# References


- Sennrich R, Haddow B, Birch A. “**Improving Neural Machine Translation Models with Monolingual Data**.” In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2016 Aug (pp. 86-96).
- Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. "**Unsupervised Neural Machine Translation.**" In *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*.
- Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "**Unsupervised Statistical Machine Translation.**" In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3632-3642. 2018.
- Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. "**Iterative back-translation for neural machine translation**." In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24. 2018.
- Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. "**An Effective Approach to Unsupervised Machine Translation.**" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 194-203. 2019.
- Lample, Guillaume, Alexis Conneau, and Ludovic Denoyer. "**Unsupervised Machine Translation Using Monolingual Corpora Only.**" In *Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018)*.
- Lample, Guillaume, Myle Ott, Alexis Conneau, and Ludovic Denoyer. "**Phrase-Based & Neural Unsupervised Machine Translation.**" In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5039-5049. 2018.

# References


- Yang, Zhen, Wei Chen, Feng Wang, and Bo Xu. "**Unsupervised Neural Machine Translation with Weight Sharing.**" In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 46-55. 2018.
- Wu, Jiawei, Xin Wang, and William Yang Wang. "**Extract and Edit: An Alternative to Back-Translation for Unsupervised Neural Machine Translation.**" 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), pp. 1173-1183. 2019.
- Ren, Shuo, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. "**Unsupervised Neural Machine Translation with SMT as Posterior Regularization.**" (2019). In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 241-248. 2019.
- Conneau, Alexis, and Guillaume Lample. "**Cross-lingual language model pretraining.**" In *Advances in Neural Information Processing Systems*, pp. 7059-7069. 2019.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. "**MASS: Masked sequence to sequence pre-training for language generation**." In *International Conference on Machine Learning*, pp. 5926-5936. 2019.
- Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou and Shuai Ma. "**Explicit Cross-lingual Pre-training for Unsupervised Machine Translation**".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)*, pp. 770-779. 2019.

# References


- Graça, Yunsu Kim Miguel, and Hermann Ney. "**When and Why is Unsupervised Neural Machine Translation Useless?.**" In *22nd Annual Conference of the European Association for Machine Translation*, p. 35.
- Marchisio, Kelly, Kevin Duh, and Philipp Koehn. **"When Does Unsupervised Machine Translation Work?.**" *arXiv preprint arXiv:2004.05516* (2020).
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "**BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.**" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2020.
- Banerjee Tamali, Murthy V Rudra, Bhattacharyya Pushpak. **Ordering Matters: Word Ordering Aware Unsupervised NMT**. arXiv preprint arXiv:1911.01212. (2019).
- Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. "**Word translation without parallel data.**" In *International Conference on Learning Representations*. 2018.
- Nima Pourdamghani and Kevin Knight. 2017. "**Deciphering Related Languages.**" In *Proceedings of EMNLP 2017*
- Iftekhar Naim, Parker Riley, and Daniel Gildea. 2018. "**Feature-Based Decipherment for Machine Translation. *Computational Linguistics*.**"
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). **Natural language processing (almost) from scratch**. Journal of Machine Learning Research
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). **A neural probabilistic language model**. Journal of Machine Learning Research

# References


- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). **Distributed representations of words and phrases and their compositionality**. Advances in Neural Information Processing Systems 26 (NIPS 2013)
- S Chandar AP, S Lauly, H Larochelle, M Khapra, B Ravindran, VC Raykar, **An autoencoder approach to learning bilingual word representations**. Advances in neural information processing systems 27, 1853-1861
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** In *NAACL-HLT (1)*. 2019.
- Kakwani, Divyanshu, Anoop Kunchukuttan, Satish Golla, N. C. Gokul, Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. **iNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4948-4961. 2020.
- Jawanpuria, Pratik, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. **Learning multilingual word embeddings in latent metric space: a geometric approach.** *Transactions of the Association for Computational Linguistics* 7 (2019): 107-120.
- Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. "**Iterative back-translation for neural machine translation.**" In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18-24. 2018.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. "**Improving Neural Machine Translation Models with Monolingual Data.**" In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86-96. 2016.

# References


- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. **Multilingual denoising pre-training for neural machine translation**. arXiv preprint arXiv:2001.08210, 2020.

# Backup Slides

# TLM

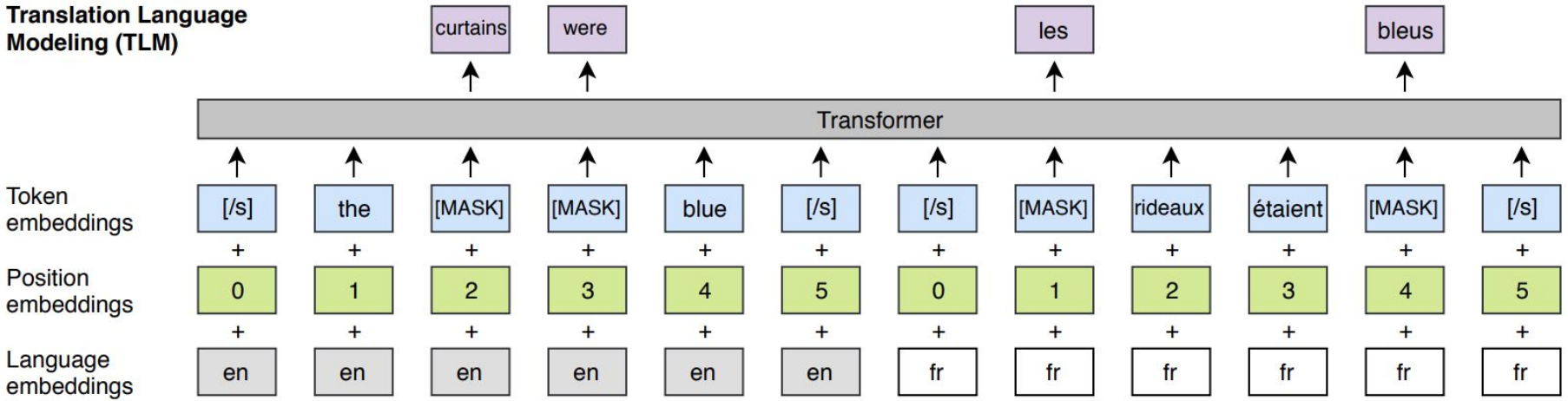
Translation Language Modelling

Cross-lingual Language Model Pretraining, ICLR, *Conneau et.al 2019*

# TLM

- XLM objective uses monolingual corpora in all the languages considered
- Does XLM learn better multilingual representations?
  - XLM objective cannot take advantage of parallel corpora if available
  - XLM objective alone cannot guarantee that the model learns better multilingual representations

# TLM (Conneau et.al 2019 )



# TLM (Conneau et.al 2019 )

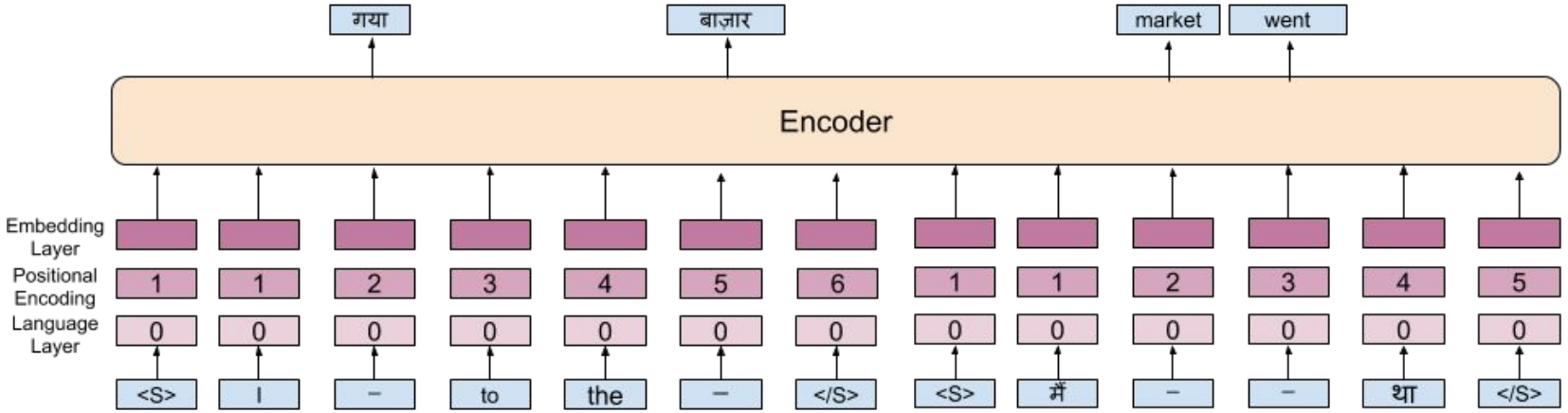
- In addition to access to monolingual corpus, we assume access to parallel corpus
- Given a parallel sentence,
  - The two sentences are concatenated and a special sentence delimiter is added to differentiate the two sentences
  - The positional information is reset to start from zero for the second language
  - The model can look at information from the context of either of the languages to predict the missing word

# TLM (Conneau et.al 2019 ) : XNLI Results

| | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Machine translation baselines (TRANSLATE-TRAIN)* | | | | | | | | | | | | | | | | |
| Devlin et al. [14] | 81.9 | - | 77.8 | 75.9 | - | - | - | - | 70.7 | - | - | 76.6 | - | - | 61.6 | - |
| XLM (MLM+TLM) | 85.0 | 80.2 | 80.8 | 80.3 | 78.1 | 79.3 | 78.1 | 74.7 | 76.5 | 76.6 | 75.5 | 78.6 | 72.3 | 70.9 | 63.2 | 76.7 |
| *Machine translation baselines (TRANSLATE-TEST)* | | | | | | | | | | | | | | | | |
| Devlin et al. [14] | 81.4 | - | 74.9 | 74.4 | - | - | - | - | 70.4 | - | - | 70.1 | - | - | 62.1 | - |
| XLM (MLM+TLM) | 85.0 | 79.0 | 79.5 | 78.1 | 77.8 | 77.6 | 75.5 | 73.7 | 73.7 | 70.8 | 70.4 | 73.6 | 69.0 | 64.7 | 65.1 | 74.2 |
| *Evaluation of cross-lingual sentence encoders* | | | | | | | | | | | | | | | | |
| Conneau et al. [12] | 73.7 | 67.7 | 68.7 | 67.7 | 68.9 | 67.9 | 65.4 | 64.2 | 64.8 | 66.4 | 64.1 | 65.8 | 64.1 | 55.7 | 58.4 | 65.6 |
| Devlin et al. [14] | 81.4 | - | 74.3 | 70.5 | - | - | - | - | 62.1 | - | - | 63.8 | - | - | 58.3 | - |
| Artetxe and Schwenk [4] | 73.9 | 71.9 | 72.9 | 72.6 | 73.1 | 74.2 | 71.5 | 69.7 | 71.4 | 72.0 | 69.2 | 71.4 | 65.5 | 62.2 | 61.0 | 70.2 |
| XLM (MLM) | 83.2 | 76.5 | 76.3 | 74.2 | 73.1 | 74.0 | 73.1 | 67.8 | 68.5 | 71.2 | 69.2 | 71.9 | 65.7 | 64.6 | 63.4 | 71.5 |
| XLM (MLM+TLM) | **85.0** | **78.7** | **78.9** | **77.8** | **76.6** | **77.4** | **75.3** | **72.5** | **73.1** | **76.1** | **73.2** | **76.5** | **69.6** | **68.4** | **67.3** | **75.1** |

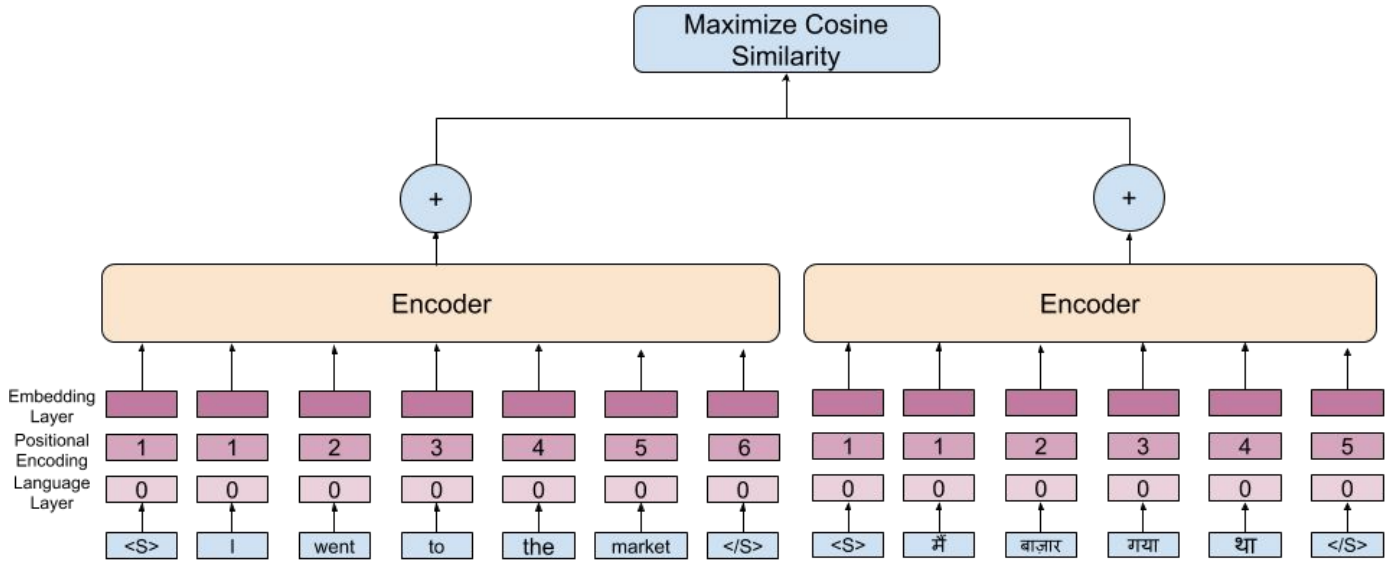
# Extensions to TLM

- TLM model does not fully utilize the potential of parallel corpus
- Modify TLM objective to predict aligned words from the other language



# Extensions to TLM

- Maximize the cosine similarity between the encoder representation of the two sentences



# Challenges in Indic Languages?

| Original Sentence | Comments | Google Translate[30 Nov,2020] |
|---|---|---|
| ನಾನು ಹೇಳುವುದನ್ನು ಸರಿಯಾಗಿ ಕೇಳಿಸಿಕೋ<br>nAnu heLuvudannu sariyAgi keLisiko<br>Me telling correctly listen | Literary Language | Listen to me correctly |
| ನಾನು ಹೇಳೋದನ್ನ ಸರ್ಯಾಗಿ ಕೇಳ್ಸೊ<br>nAnu heLodanna saryAgi keLsko | Spoken Language | I am Sergio Katsko of Noodon |
| ಊಟ ಮಾಡಿಕೊಂಡು ಹೋಗು<br>UTa mADikoMDu hogu<br>Lunch have go | Literary Language | Go Have lunch<br>(Go after having lunch) |
| ಊಟ ಮಾಡ್ಕೊಂಡು ಹೋಗು<br>UTa mADkoMDu hogu | Spoken Language | Modify the meal |

Phenomenon similar to Schwa Deletion in Literary language and Spoken language

# Why Indic Languages?

| Original Sentence | Comments | Google Translate |
|---|---|---|
| ವಂಚಕಾಸುರರನ್ನೊದ್ದೋಡಿಸಿರುವವನಾರೆಂದೇನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ? | Maximum Sandhi transformation | Do you know anyone who has cheated? |
| ವಂಚಕ ಅಸುರರನ್ನು ಒದ್ದು ಓಡಿಸಿ ಇರುವ ಅವನು ಯಾರು ಎಂದು ಏನು ಆದರು ನಿಮಗೆ ತಿಳಿದು ಇದೆಯೇ ? | No Sandhi transformation | Do you know who became the one who drove out the crafty demons? |
| ವಂಚಕಾಸುರರನ್ನು ಒದ್ದೋಡಿಸಿರುವವನು ಯಾರೆಂದು ಏನಾದರು ನಿಮಗೆ ತಿಳಿದಿದೆಯೇ ?<br><br>Crooked demons one who kicked them away who is you know | Normal Usage | Do you know who is the one who kicked the crooks? |

# Components of U-MT

- Suitable initialization of the translation models: This helps the model to jump-start the process.
- Language modeling: This helps the model to encode and generate sentences.
- Iterative back-translation: It bridges the gap between encoder representation of a word in source and target languages.

# Adding subword information

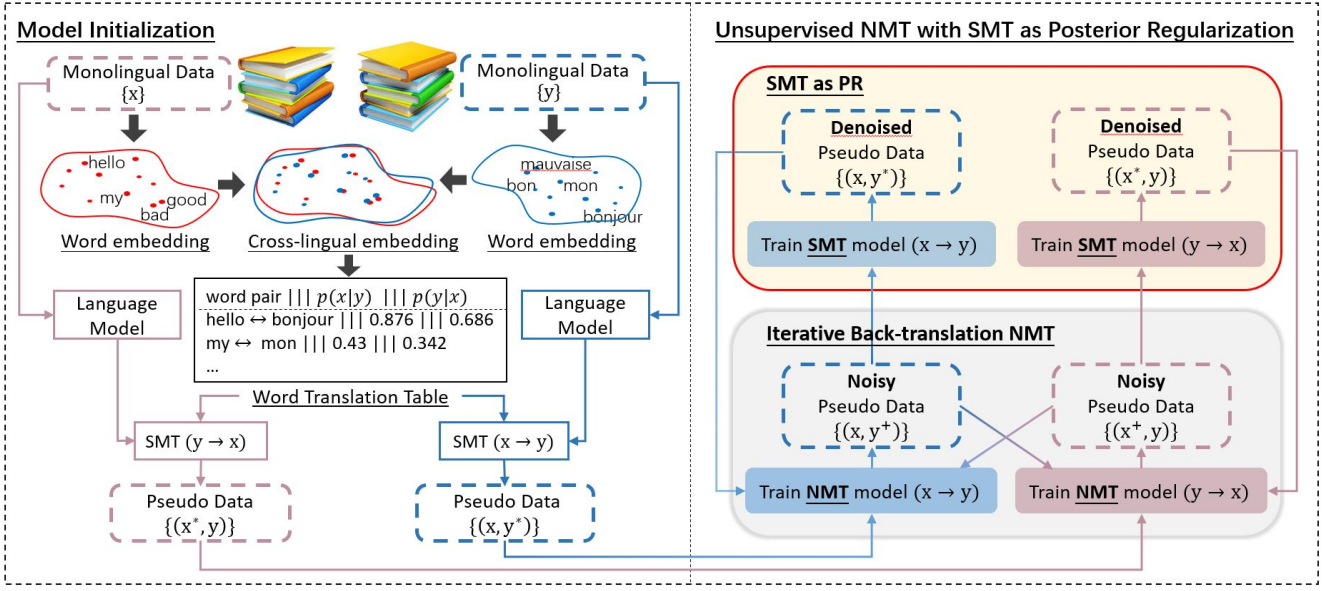
- We want to favor translation candidates that are similar at the character level.
- Additional weights are added to initial phrase-table like lexical weightings.
    - Unlike lexical weightings it use **a character-level similarity function** instead of word translation probabilities.

$$\text{score}(\bar{f}|\bar{e}) = \prod_i \max\left(\epsilon, \max_j \text{sim}(\bar{f}_i, \bar{e}_j)\right)$$

$$\text{sim}(f, e) = 1 - \frac{\text{lev}(f, e)}{\max(\text{len}(f), \text{len}(e))}$$

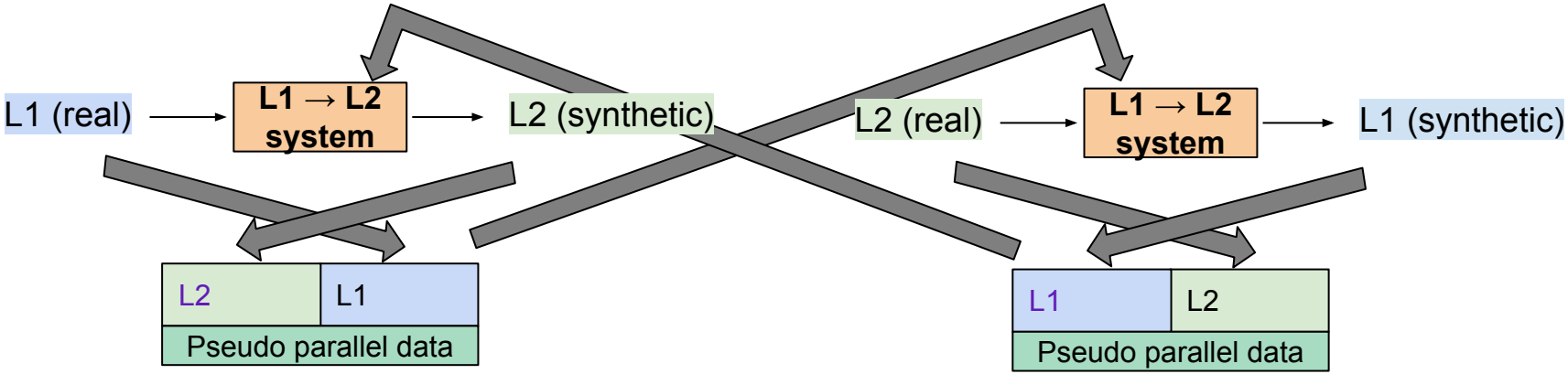
# USMT as Posterior Regularization

- USMT initialisation.
- UNMT backtranslation training with SMT as Posterior Regularization.
  - Posterior Regularization: An SMT system to filter out noises using phrase table. It eliminates the infrequent and bad patterns generated in the back-translation iterations of NMT



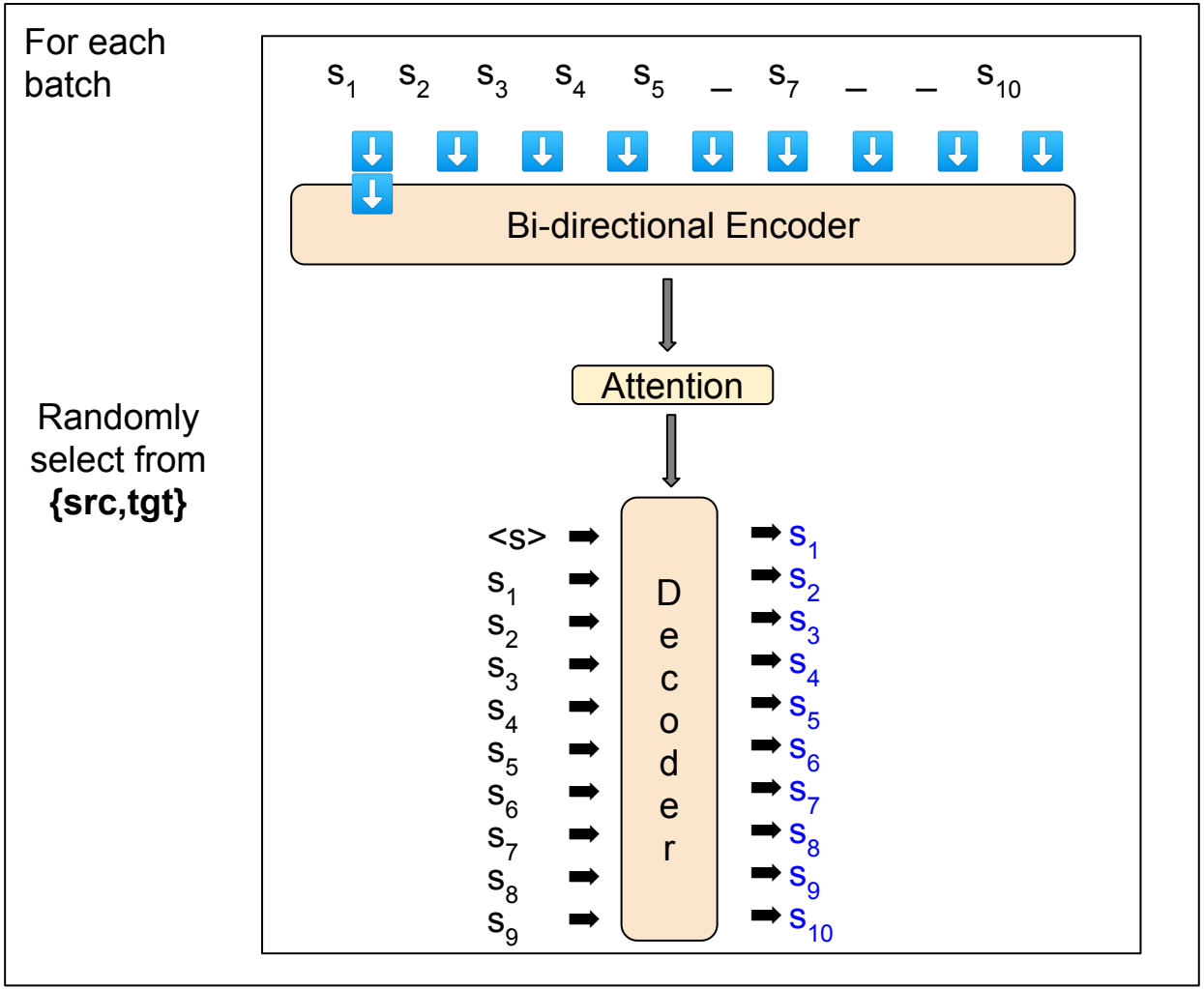
# Iterative refinement

- Generate a synthetic parallel corpus by translating the monolingual corpus with the initial system L1→L2, and train and tune SMT system L2→L1.
  - To accelerate our experiments, use a random subset of 2 million sentences from each monolingual corpus for training.
  - Reuse the original language model, which is trained in the full corpus.
- The process is repeated iteratively until some convergence criterion is met.



# BART Pretraining

- Trained by
  - Corrupting text with an arbitrary noising function
  - Learning a model to reconstruct the original text.
- Denoising full text

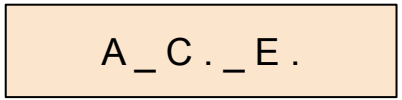



Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL 2020)*

# BART pretraining (noising steps) (Lewis et al. 2020)

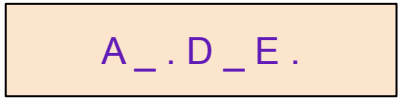

Original document

Token Masking

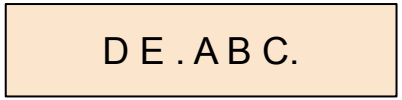



Token deletion


Text infilling




Sentence permutation




Document rotation