# Infosys Machine Translation System for WMT20 Similar Language Translation Task

**Kamalkumar Rathinasamy, Amanpreet Singh, Balaguru Sivasambagupta,**
**Prajna Prasad Neerchal, Vani Sivasankaran**
Infosys Limited
{kamalkumar_r,amanpreet.singh04,balaguru.s,prajna.neerchal,vani.s}@infosys.com

## Abstract

This paper describes Infosys' submission to the WMT20 Similar Language Translation shared task. We participated in Indo-Aryan language pair in the language direction Hindi to Marathi. Our baseline system is byte-pair encoding based transformer model trained with the fairseq sequence modeling toolkit. Our final system is an ensemble of two transformer models, which ranked first in the WMT20 evaluation. One model is designed to learn the nuances of translation of this low resource language pair by taking advantage of the fact that the source and target languages are the same alphabet languages. The other model is the result of experimentation with the proportion of back-translated data to the parallel data to improve translation fluency.

## 1 Introduction

Neural Machine Translation (Bahdanau et al., 2015; Vaswani et al., 2017) is the most popular approach for machine translation. Transformer-based NMT has outperformed many recurrent neural network based models. There is scope for improvement in NMT, particularly for low-resource language pairs.

Our techniques are experimented on the fairseq sequence modeling toolkit (Ott et al., 2019) for NMT. Our system is an ensemble of two transformer-based models. One designed for low-resource language pairs by taking advantage that both are same alphabet languages. The other model is built after experimenting on renowned back-translation technique (Sennrich et al., 2016a) by exploiting target monolingual data.

## 2 Data

Hindi-Marathi bitext data contains ∼49K sentence pairs. Target monolingual data comprises of 326K Newscrawl sentences and 10,839K raw sentences.

### 2.1 Data Preprocessing

Typical training sentence pairs comprises of a source and a target sentence. There are ∼1K training sentence pairs where source or target contains multiple sentences delimited by '/'. Matching pair for these sentences is derived based on the proximity of token lengths between source and target sentence.

Non-printable characters are removed, punctuations are normalized, and the data is tokenized, with the Moses tokenizer. Byte-pair encoding (BPE) has been adopted (Sennrich et al., 2016a) to build source and target sub-word vocabularies of size 22.5K and 32.8K respectively, when configured to construct with 60K symbols.

### 2.2 Data filtering

#### 2.2.1 Bitext data

Sentences with more than 175 words, sentences with no words, and sentence pairs exceeding length ratio of 1.5 are removed from training data. This eliminated around 18% of the overall real bitext data.

#### 2.2.2 Synthetic data

CommonCrawl n-grams raw monolingual files are processed[1] to remove sentences with invalid characters, strip leading and trailing whitespaces, and remove duplicate sentences.

## 3 System Overview

Our Hindi-Marathi primary system is an ensemble of two transformer models. One is back-translated model and the other model is trained on anonymized data.

[1] https://github.com/kpu/preprocess

*Proceedings of the 5th Conference on Machine Translation (WMT)*, pages 437–441
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

### 3.1 Base Model Architecture and Hyperparameters

Our model is built using fairseq[2] (Ott et al., 2019) toolkit. The Transformer, an encoder-decoder architecture (Vaswani et al., 2017), with 6 layers for the encoder and 6 layers for the decoder, and with 8 heads in all multi-head attention layers, is our base model. Embedding dimension is set to 512 and feed-forward size (FFN) is set to 2048. Our model is trained on single GPU with maximum tokens per GPU set to 4096. The batch size multiplier is set to 8. Dropout probability of 0.3 and label smoothing probability of 0.1 is applied to avoid overfitting. Adam optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.98$. The model is trained with an initial learning rate of 5e-4 and 4000 warm-up updates. The ensemble model prepared by averaging last 3 checkpoints is used for inferencing. Reported detokenized test BLEU is 9.13 for the provided dev dataset.

Parameters are tuned, following Baquero-Arnal et al. (2019). Threshold frequency is set such that only tokens occurring at least 10 times in the training data will be part of the vocabulary. Maximum tokens per GPU is set to 4000 and the batch size multiplier is fixated to 4 to set an effective batch size of 16000 tokens with dropout probability of 0.1. This led to improved performance. Reported detokenized test BLEU is 14.13 and hence these settings are adapted.

### 3.2 Backtranslation

Back-translation is a popularly adapted data augmentation technique which aids in building better NMT systems, especially for low resource language pairs by leveraging monolingual corpora (Sennrich et al., 2016a). An intermediate system is first trained on parallel data which is used to translate target monolingual data into source language. Sampling is used as a method for inference (Edunov et al., 2018). Synthetic parallel data is constructed from the intermediate system generated synthetic source while the target is the provided monolingual data. The Bitext data filters are also applied to synthetic data but only removed sentences with more than 250 words. New training data is constructed by appending this synthetic parallel data to real bitext data and a final system that will translate from the source to the target language will be trained.

#### 3.2.1 Bitext and Synthetic corpora proportion

**Related Work** Real to synthetic parallel data close to 1-to-1 proportion works best for Sennrich et al. (2016a). Junczys-Dowmunt et al. (2016), also chose 1-to-1 ratio of real to synthetic parallel data for English-Russian news translation task. It is also known from past experiments that increasing the ratio of synthetic training data erratically, degrades system performance, depending on quality and domain of synthetic data (Sennrich et al., 2016a; Currey et al., 2017; Poncelas et al., 2018).

In contrast, experiments conducted by Stahlberg et al. (2018), shows that performance of system does not reduce as long as the ratio of real parallel to synthetic parallel data does not exceed 1-to-8 (1.6M out of 3M Turkish monolingual data is preferred for training along with 0.2M of parallel corpus for English-Turkish). Fadaee and Monz (2018), claims, 1-to-5 real to synthetic parallel data ratio achieved best performance in news translation task for German-English with 4.5M parallel corpus.

This limits from taking advantage of all available monolingual corpus. Only a small portion of it can be used as synthetic parallel training data. Oversampling (Chu et al., 2017; Junczys-Dowmunt and Grundkiewicz, 2018) real parallel data can overcome this problem. By oversampling primary parallel data equivalent to the synthetic parallel data from all monolingual data, effective 1-to-1 ratio of bitext and synthetic parallel data can be retained.

**Experiment** 1-to-1 ratio of bitext to synthetic data is chosen after experimentation with ratios (see Table 1, Figure 1).

| Ratio | BLEU |
|---|---|
| Baseline (1:0) | 14.13 |
| 1:0.5 | 18.08 |
| 1:1.0 | 18.76 |
| 1:2.5 | 16.20 |
| 1:5.0 | 14.49 |
| All monolingual data (1:78.0) | 11.01 |

Table 1: BLEU score for different bitext and synthetic corpora proportion

It is crucial to find the ideal proportion of synthetic data to use. Utilization of all available out-of-domain and raw monolingual corpora to the maximum effect can be further explored.

[2] https://github.com/pytorch/fairseq

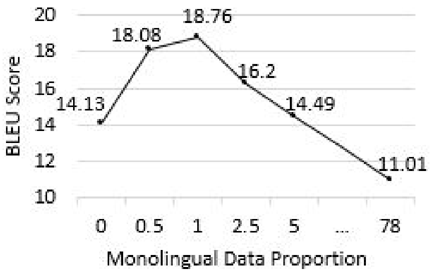

Figure 1: BLEU score for different bitext and synthetic corpora proportion

### 3.2.2 Out-of-domain data

**Handling OOV** BPE is applied on monolingual target data using byte pairs learnt during bitext BPE operation. Out-of-vocabulary (OOV) tokens in BPE applied monolingual target data are the tokens not in bitext vocabulary. These out-of-vocabulary tokens are replaced by a special symbol UNK in the monolingual target data (see Table 2).

| Considered | Filtered | OOV |
|---:|---:|---:|
| 25K | 19K | 16.5% |
| 50K | 35K | 17.9% |
| 100K | 72K | 22.0% |
| 300K | 172K | 21.2% |
| 11.2M | 2.5M | 23.0% |

Table 2: Out-of-domain words in target monolingual data. ”Considered” represents the amount of target monolingual data used for study. ”Filtered” represents the amount of target monolingual data after applying filters.

**Experiment** Since the intermediate model spot UNK symbol in the inputs during inferencing, inferenced data also contains UNK symbol.

Gulcehre et al. (2015), claims to eliminate monolingual sentences with more than 10% UNK symbols for better performance. Sennrich et al. (2016b), claims to handle rare/unseen words by representing it in a sequence of sub-word units using existing vocabulary that was learnt on the parallel data. Our systems are experimented by excluding sentences with UNK symbol.

Systems are trained with different proportions of real to synthetic data by eliminating all sentence pairs containing UNK in training data. Table 3 shows the study of model performance before and after removing sentence pairs containing UNK. 1-to-1 proportion of real and synthetic data with out-of-vocabulary tokens masked by UNK symbol scored best (18.76) out of all outcomes.

| Ratio | All Data | Data without UNK |
|---|---|---:|
| 1:0.5 | 18.08 | 17.73 |
| 1:1.0 | 18.76 | 18.34 |
| 1:5.0 | 14.49 | 16.60 |

Table 3: BLEU scores on models with and without removing sentences containing UNK

### 3.3 Anonymization

Analysis of the results of the model achieved 18.76 BLEU score, reveals that the translation accuracy is negatively impacted when UNK is generated. This is handled by building another model with bitext data only, where the similarity between source and target languages are anonymized by masking. This approach enables the model to specifically focus on learning the nuances of translation only (i.e.., enables the model to focus on the specific section in the source sentence that gets altered during translation).

Language pair comprising same alphabetic languages contains same words between them carrying similar meaning. Numbers, names, geographic names, etc., also holds same script. i.e. tokens that are not language specific. The approach here is to anonymize those words which are equally present in source and target sentences. One special character is used to mask all those tokens. The special character is chosen in place of a special word to eliminate the possibility of splitting the special word during sub-word tokenization.

This approach reduces the vocabulary size and the learning parameters of the model, preserving the context. This results in transforming sentences which appeared to be different in its raw form into duplicate sentences in its anonymized form, which are then deduplicated.

Hi-Mr track with ∼49K training sentences without masking technique generated source and target vocabulary of size 22.5K and 32.8K respectively. Anonymization reduced source and target vocabulary size to 20.9K and 31.0K respectively. This approach resulted in improvement of BLEU score by 1.2 over baseline. The impact of this approach is proportional to the similarity of source and target languages. The key observation is that this model performed better at translation of sentences that are translated poorly (with UNK tokens) by back-translation model.

### 3.4 Stacking

Benefits of both the masking systems (masking OOV tokens with UNK symbol and masking similar tokens) are attained through stacking. Model trained on anonymized parallel data and the model trained on real bitext plus synthetic parallel data are ensembled to achieve 19.76 BLEU with Dev data.

### 3.5 Post-processing

The anonymized words are preserved before inferencing and the inference results are decoded by replacing the special symbols with the preserved anonymized tokens followed by BPE detokenization.

## 4 Results

Our novel anonymization technique improved BLEU by 1.2. Optimal proportion of back-translated data improved BLEU by 3.5. Ensembling best systems improved BLEU by 1.0. (See Table 4)

| **System** | **Dev BLEU** |
|---|---:|
| Baseline | 9.13 |
| +hyperparameter tuning | 14.13 |
| +anonymization | 15.46 |
| Baseline | 9.13 |
| +hyperparameter tuning | 14.13 |
| +backtranslation | 18.76 |
| Ensemble | 19.71 |

Table 4: BLEU scores on Hindi-Marathi

Our final submission to the competition in Hindi-Marathi track achieved 18.26 BLEU and ranked first among all submissions.

## 5 Conclusion

This paper describes the techniques involved in our system submitted for the WMT20 Similar Language Translation task by Infosys. This winning Hindi-Marathi translation system is built based on NMT and evaluated based on the metric, BLEU.

The domain-based data preprocessing and filtering techniques eases model learning. Adopting novel approach of anonymizing language agnostic tokens aided our system to focus more on tokens that matters in the translation. It is highly observed that the ratio of monolingual data used against bitext data plays a vital role in back-translated models. Improving translation accuracy and language fluency by utilizing all available out-of-domain monolingual corpora to the maximum effect can be further explored.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*, San Diego, USA.

Pau Baquero-Arnal, Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. 2019. The mllp-upv spanish-portuguese and portuguese-spanish machine translation systems for wmt19 similar language translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 179–184, Florence, Italy.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, Vancouver, Canada.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. pages 489–500.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016. The amu-uedin submission to the wmt16 news translation task: Attention-based nmt models as feature functions in phrase-based smt. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. Ms-uedin submission to the wmt2018 ape shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, USA.

A Poncelas, D Shterionov, A Way, GM de Buy Wenniger, and P Passban. 2018. Investigating back-translation in neural machine translation. arxiv 2018. *CoRR*, abs/1804.06189.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany.

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. pages 204–211.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.