# Multilingual NLP 11-737 – Assignment 2: Multilingual Translation

**Kushagra Mahajan, Nikhil Gupta, Shubham Phal**
School of Computer Science
Carnegie Mellon University
kmahajan@andrew.cmu.edu, nikhilgu@andrew.cmu.edu, sphal@andrew.cmu.edu

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 1.71 | -1.1919 |
| eng-aze | 1.47 | -1.3069 |
| bel-eng | 1.34 | -1.3722 |
| eng-bel | 1.20 | -1.3988 |

Table 1: Bilingual Training Baseline

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 11.93 | -0.2240 |
| eng-aze | 6.02 | -0.0898 |
| bel-eng | 17.41 | -0.3295 |
| eng-bel | 9.76 | -0.4137 |

Table 2: Multilingual Training Baseline

## Abstract

The goal of neural machine translation is to predict a sentence in target language given the source language. In this work we explore neural machine translation in extreme low resource settings. We first evaluate the baseline methods and then experiment with multiple techniques to improve the multilingual transfer namely backtranslation, cross-lingual transfer and better modelling. We summarize our findings for each case.

.

## 1 Baseline Analysis

The results of bilingual training are shown in Table 1. The results of multilingual training are shown in Table 2. The results of finetuning flores-101 are shown in Table 3.

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 12.98 | -0.0940 |
| eng-aze | 8.04 | 0.0323 |
| bel-eng | 20.26 | -0.0306 |
| eng-bel | 17.69 | 0.1696 |

Table 3: Finetuning Pretrained Multilingual Model Baseline

## 2 Analysis of Multilingual Pretraining

Based upon our analysis of the baseline methods we find that multilingual pretraining positively impacts the quality of translations. This is evident from the large increase in BLEU and COMET scores over bilingual training. In general we observe that the performance is better in translating a low resource language to English than vice versa. We hypothesize that this could be due to English being relatively easier to translate to due it simpler morphology(less agglutinating) and relatively fixed word order. On the contrary Azerbaijani and Belarusian are highly agglutinating languages with relatively free word order which leads them to have a larger vocabulary and makes them difficult to translate to.

## 3 Improvements

### 3.1 Backtranslation

In this experiment we evaluate the backtranlsation approach proposed by (Edunov et al., 2018). Backtranslation is a four step process which involves

1. Retrieve the monolingual data (MD).
2. Binarize and backtranslate MD.
3. Augment the gold standard parallel data (AD).
4. Binarize and retrain the model over AD

We perform different experiments and arrive at new insights. Given the budget and the time constraints of the project we limit our evaluation to measuring the performance of backtranslation over Azerbaijani(aze) to English(en) and vice versa. We start with a large monolingual aze corpus CC-100-aze (Conneau et al., 2020) that contains about 41 million( 7GB) aze sentences of news commentary data.

In the first experiment we perform simple bilingual training with fairseq and use it for backtranslating the first 50K sentences(for the purpose of

reproducibility) from the dataset. Consistent with the observations of (Edunov et al., 2018) we find that using fairseq with a greedy decoding strategy yields extremely noisy translations (verified by feeding the first few translations into google translate). Hence we experiment with beam search and top-k sampling. With beam search (k=5) we observe that the translation quality continues to be poor. Additionally we observe that in multiple cases the backtranslation model keeps repeating words in the output. Nevertheless we use this backtranslated data as is, to create new parallel data and retrain our bilingual model on this augmented training set. We observe that for aze-eng translation task our BLEU score drops to 0.94 while COMET drops to -1.3224. We hypothesize that this could be a result of the bad translations that were added to the parallel data. To test this hypothesis we remove the bad translations from our previous output resulting in 15547 sentences. In this case our criteria for removing bad translations simply involves removing those translations for which a token was repeated more than three times. Contrary to our hypothesis we find that retraining our model with this new parallel data causes our BLEU scores to drop further to 0.22 and COMET score to drop to -1.3302. We hypothesize that the poor translations could be due to

1. The selected monolingual data (news commentary) being completely out of domain with the training data(ted talks) i.e very little vocabulary overlap.

2. A weak translation model

To solve the first problem we introduce a simple preprocessing step on the monolingual data. We observe that the average length of an aze sentence in the gold standard training corpus is about 12. We then construct a vocabulary for the aze data by finding the 10000 most frequently occurring tokens. We then iterate over the first 200000 examples from CC-100-aze dataset and retain only those sentences which contain atleast 9 tokens present in the vocabulary and have a length less than 16. We find 6920 such examples in the dataset. To address the second problem we perform multilingual pretraining over aze and turkish and use it for back translation. In line with our expectations we find that this setup improves our BLEU and COMET scores over the baseline as shown in Table 4. We further check if

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 2.82 | -1.0083 |
| eng-aze | 2.32 | -1.0667 |

Table 4: Bilingual Training with Backtranslation Performance

adding additional monolingual data improves backtranslation performance. Specifically we apply the same steps to the first 500000 examples from CC-100-aze and find 17256 qualified examples. We observe that in this setup our BLEU score drops to 0.25 and COMET score drops to -1.9501. Based on this observations we hypothesize that having too much of synthetic data is detrimental to the performance of back translation. This again is consistent with the views of (Conneau et al., 2020) In our final experiment we evaluate backtranslations with top10 sampling. Contrary to the findings of (Conneau et al., 2020) we find that top10 sampling performs worse than beam search with our BLEU scores dropping to 0.35 and COMET scores dropping to -1.9946. We hypothesize that this could be due to two reasons. By applying preprocessing we force the monolingual data to be in domain with the available parallel corpus. Thus in this setting the diverse outputs generated by top10 sampling perform worse than the most likely output generated by beam search. The other reason could be that unlike (Conneau et al., 2020) who performed their analysis on a relatively large german corpus we perform this analysis in an extremely low resource aze corpus with just 5946 training examples

### 3.2 Multilingual Pre-trained Language Model

To compare and assess the effectiveness of different modelling approaches we experiment with different state-of-the-art multilingual pre-trained language models namely MBart-50 (Tang et al., 2020), and M2M-100 (Fan et al., 2021). We use these pretrained models off the shelf to assess the quality of translations on Azerbaijani, Belarusian and English. As MBart-50 was not trained on Belarusian (bel), we use it only to translate Azerbaijani to English and vice versa. The results obtained are thought provoking. An off-the-shelf MBart-50 model outperforms all our baselines methods. We hypothesize that this superior performance could be a result of extensive cross-lingual transfer taking place in MBart-50. For the M2M-100 model we find that training and evaluation is much faster than MBart-50 but the translations are of poorer

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 17.44 | 0.1795 |
| eng-aze | 8.13 | 0.2421 |

Table 5: MBart-50 Pretrained Performance

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 8.51 | -0.4084 |
| eng-aze | 3.41 | -0.1575 |
| bel-eng | 12.64 | -0.4713 |
| eng-bel | 5.42 | -0.7610 |

Table 6: M2M-100 Pretrained Performance

quality as shown in table 6. The results of M2M-100 are still better than our bilingual baseline approach. The results highlight the importance of pretraining on large-scale monolingual corpora in many languages for improved downstream tasks like supervised and unsupervised machine translation. Though such extensive multilingual models give performance boost for all languages, particularly significant gains are seen for low-resource languages. The pretrained model allows us to better model the multilingual translation task, and finetuning can lead to even greater improvements. Our hypothesis for M2M-100 performing poorly compared to MBart-50 without finetuning is that M2M-100 is trained for a much larger set of languages and a much larger set of language pair tasks. Hence, it is quite likely that its performance would surpass the MBart-50 performance on finetuning. However, the performance of the direct pretrained model for M2M-100 of is lower than MBart50. This hypothesis is validated by the authors' findings in (Fan et al., 2021).

### 3.3 Cross-Lingual Transfer

In this method of improving low-resource machine translation, we start by choosing a high-resource transfer language which is similar to the low-resource language. The training is first done for the high-resource language pair. The model weights are then used to initialize the training for the language pair containing the low-resource language. We use this method to improve the performance from Azerbaijani to English. (Lin et al., 2019) shows a good transfer language for Azerbaijani is Turkish due to their similarity. We begin by first fine-tuning the small FLORES-101 model for Turkish to English translation. Then we take these model weights and use it to initialize training for Azerbaijani to English translation. After convergence, we see an increase in the BLUE score by 1 (13.95) as compared to the fine-tuning baseline. We observe that this score is more than the Multilingual Training baseline score (where Aze and Tur text was used together to train a single model at once). We hypothesize that this is because in the latter case the model has to retain more information (with the same model capacity) during training to perform good on both Aze-Eng and Tur-Eng translations as the loss function penalizes both but in our approach, the model can just be optimized for Aze-Eng translation.

| LP | BLEU | COMET |
|---|---|---|
| aze-eng | 13.95 | -0.1130 |
| eng-aze | 7.20 | 0.0004 |

Table 7: Cross-lingual Transfer Performance

We also applied the same technique in the reverse direction (initialized the weights of Eng-Aze training from Eng-Tur finetuning) but that resulted in a drop in BLEU score. All the results are shown in Table 7.

### 3.4 Conclusion

The report highlights the various popular methods for multilingual machine translation. We reproduced the baseline results and analyzed multiple ways to improve performance over the baselines. Most of our methods led to improvements over the baselines and validated our hypotheses. Overall, the assignment was a great learning curve for us, and gave us an opportunity to explore the state-of-the-art multilingual translation models and techniques.

## References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junx-

ian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.