# Advancing Low-Resource Neural Machine Translation

Irina Saparina, Elena Kosheleva, Taras Khakhulin, Antov Kiselev, Emil Magerramov

DeepHack.Babel, Winter 2018

## Objectives

The objective of the hackathon was to create a translation model for an unknown pair of languages. :

- A small parallel corpus and two large monolingual corpora were available.
- The solutions were tested remotely, so that the computational resources were given equally for everyone and contestants were unable to perform manual labeling.
- Only 8 hours were given for training and inference.

## Introduction

Large-scale parallel corpora are indispensable to train highly accurate machine translators. However, manually constructed large-scale parallel corpora are not freely available in many language pairs.

While neural machine translation (NMT) is making good progress in the past two years, tens of millions of bilingual sentence pairs are needed for its training. The goal of the hackathon is to explore various methods of using monolingual corpora with a small parallel corpus to relinquish the need for a huge parallel corpus for training NMT.

## Disadvantages of Unsupervised NMT

For the following reasons, it was impossible to implement a full-fledged system of unsupervised neural machine translation:

- Lack of time for training
- Small size of the provided monolingual corpora to fully capture the similarity between languages
- Absence of resources for developing system such complexity
- Complex scalability of systems

## Unsupervised models

These problems were a serious obstacle for us in solving the problem. That's why we looked at the problem from a different angle and tried to improve a **supervised** machine translation system with monolingual data.

1. Backtranslation + Denoising [1]
2. Adversarial model [2]
3. Style transfer [3]

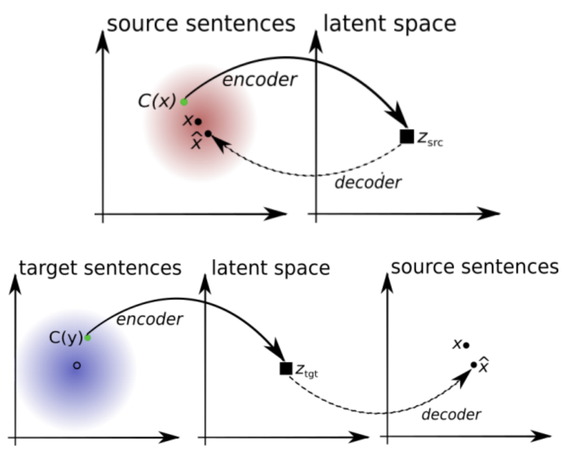


Figure 1: Adversarial model

## Methods

We tried the following method from [5]:

1. Train two models for translation in forward (source → target) and backward (target → source) directions.
2. Translate monolingual target sentences to produce synthetic source sentences.
3. Back-translate the synthetic source sentences to obtain a synthetic target sentences.
4. Calculate sentence-level similarity metric scores using the monolingual target sentences as reference and the synthetic target sentences as candidates (sent-BLEU).
5. Choose the monolingual target sentences and the corresponding synthetic source sentences with high similarity score.
6. Use the filtered synthetic source sentences as the source side and the monolingual target sentences as the target side in the additional pseudo-parallel corpus.

## Architecture

We used OpenNMT [4] framework to implement the translation model.

The experiments have shown that the method filters out too many sentences. This lead us not to use filtering and to use all of the synthetic sentences from back-translation. We tried this approach in both forward and backward directions, but it was time-consuming and we're unable to verify a BLEU score on the leader-board.

We also used the monolingual corpora to build monolingual *fastText* word embeddings.

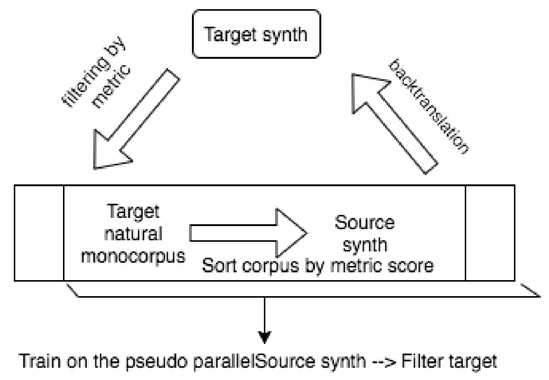


Figure 2: Final model

## Results

The approach we used got us to the second place on the first language pair and to the first place on the second language pair. The top BLEU score for the second language pair was **0.26003**.

## Conclusion

We've created an end-to-end model which translates from one unknown language to another unknown language with one of the highest scores across all teams during the competition.

## References

[1] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *CoRR*, abs/1710.11041, 2017.

[2] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017.

[3] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Style transfer from non-parallel text by cross-alignment. *CoRR*, abs/1705.09655, 2017.

[4] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

[5] Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78, 2017.

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

## Acknowledgements

We thank the organizers for an interesting problem and a warm reception.

## Contact Information

- Web: http://babel.tilda.ws
- Telegram: @irisaparina, @lkosh96, @vitaminotar, @a_kiselev, @lemhell

DeepHack.Babel

iPavlov.ai