# Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages

**Yunsu Kim**[1*], Petre Petrov[1,2*], Pavel Petrushkov[2], Shahram Khadivi[2], Hermann Ney[1]

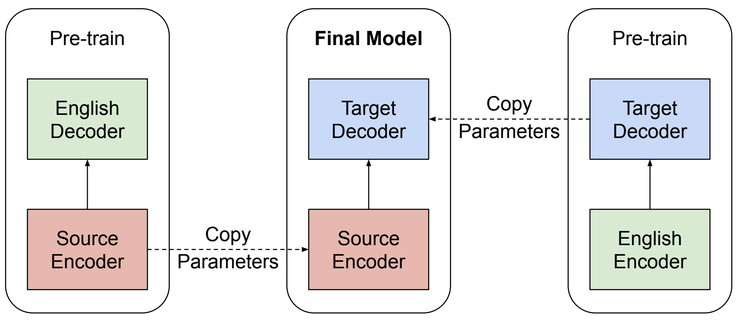
[1]Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany

[2]eBay Inc., Aachen, Germany

## Plain Transfer

Non-English language pairs: small or no parallel data (e.g. German→Czech)

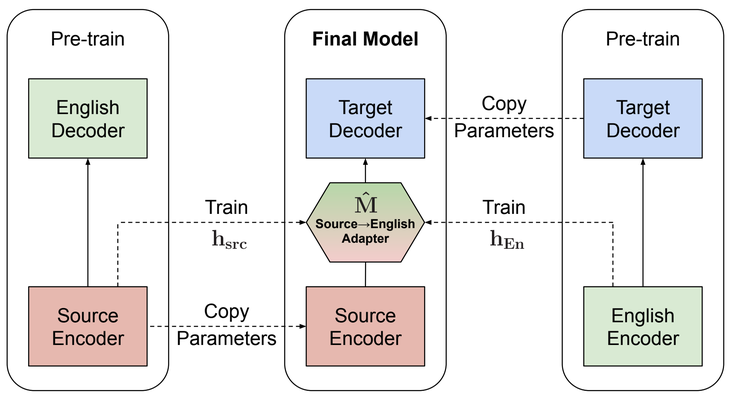
- Pre-train source→English and English→target models (English as pivot)




**Problem**: Discrepancy between pre-trained encoder and decoder

## Pivot Adapter

**Solution 1**: Insert an adapter between pre-trained encoder and decoder




Mapping $\hat{\mathbf{M}}$: source encoder output ($\mathbf{h_{src}}$) → English encoder output ($\mathbf{h_{En}}$)

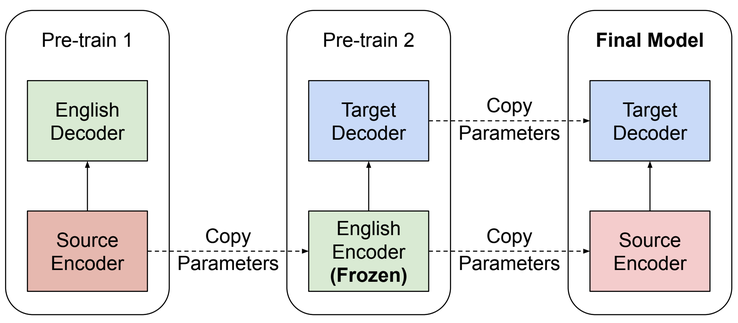
- Trained with source-English parallel data

$$\hat{\mathbf{M}} = \underset{\mathbf{M}}{\operatorname{argmin}} \sum_{(\mathbf{h_{src}}, \mathbf{h_{En}})} \|\mathbf{M}\mathbf{h_{src}} - \mathbf{h_{En}}\|^2$$

**Effect**: Make source encoder outputs compatible to target decoder

## Step-wise Pre-training

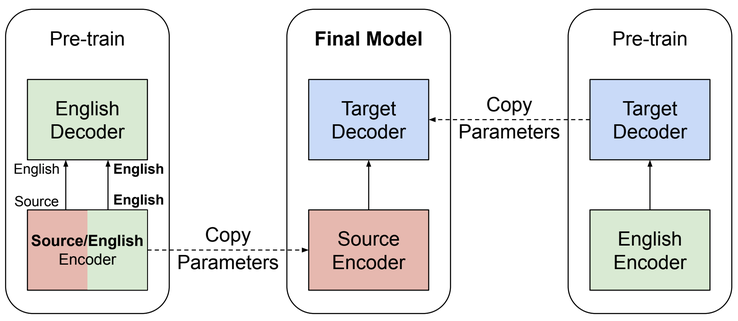
**Solution 2**: Pre-train for source→English and then English→target




**Effect**: Target decoder directly sees the source encoder space in pre-training

## Cross-lingual Encoder

**Solution 3**: Pre-train an encoder for both source language and English




Denoising autoencoder: English (noisy) → English (clean)

- Combined with the normal source→English cross entropy
- Noise: insertion, deletion, permutation

**Effect**: Encodes source/English inputs in the same space

## Zero-shot / Zero-resource Results

**Scenario**: Given **no** source-target parallel data

- Zero-shot: no further tuning of the pre-trained parameters
- Zero-resource: fine-tune the pre-trained parameters with synthetic source-target parallel data (e.g. pivot-based forward translation)

| Results in BLEU [%] | | French→German | | German→Czech | |
|---|---|---|---|---|---|
| | | test2012 | test2013 | test2012 | test2013 |
| | Pivoting (cascaded) | 16.6 | 17.9 | 16.4 | **19.5** |
| | Multilingual many-to-many | 14.1 | 14.6 | 5.9 | 6.3 |
| Zero-shot | Plain transfer | 0.1 | 0.2 | 0.1 | 0.1 |
| Zero-shot | + Pivot adapter | 0.1 | 0.1 | 0.1 | 0.2 |
| Zero-shot | Step-wise pre-training | 11.0 | 11.5 | 6.0 | 6.5 |
| Zero-shot | + Cross-lingual encoder | 17.3 | 18.0 | 14.1 | 16.5 |
| Zero-res | + Synthetic data (10M) | **19.3** | **20.9** | **16.5** | 19.1 |

**Conclusion**: Built decent NMT models without any real parallel data

- Outperforms pivoting and multilingual systems
- German→Czech test2019: **17.2** (ours) vs. 15.5 (NICT, unsupervised)

## Small-scale Fine-tuning Results

**Scenario**: Given **small** (∼250k) source-target parallel data for fine-tuning

| Results in BLEU [%] | French→German | | German→Czech | |
|---|---|---|---|---|
| | test2012 | test2013 | test2012 | test2013 |
| Direct source→target | 14.8 | 16.0 | 11.1 | 12.8 |
| Multilingual many-to-many | 18.7 | 19.5 | 14.9 | 16.5 |
| Plain transfer | 17.5 | 18.7 | 15.4 | 18.0 |
| + Pivot adapter | 18.0 | 19.1 | 15.9 | 18.7 |
| Step-wise pre-training | 18.6 | 19.9 | 15.6 | 18.1 |
| + Cross-lingual encoder | **19.5** | **20.7** | **16.2** | **19.1** |

**Conclusion**: Effective gains from pivot adapter or cross-lingual encoder

- Up to +2.6 BLEU [%] against multilingual systems

## Large-scale Fine-tuning Results

**Scenario**: Add **large** source-target synthetic data for fine-tuning

- Real + Synthetic: 12M (French→German), 5M (German→Czech)

| Results in BLEU [%] | French→German | | German→Czech | |
|---|---|---|---|---|
| | test2012 | test2013 | test2012 | test2013 |
| Direct source→target | 20.1 | 22.3 | 11.1 | 12.8 |
| + Synthetic data | 21.1 | 22.6 | 15.7 | 18.5 |
| Plain transfer | 21.8 | 23.1 | 17.6 | 20.3 |
| + Pivot adapter | 21.8 | 23.1 | **17.6** | **20.9** |
| Step-wise pre-training | 21.8 | 23.0 | 17.3 | 20.0 |
| + Cross-lingual encoder | **21.9** | **23.4** | 17.5 | 20.5 |

**Conclusion**: Transfer helps greatly even with large data for the main task

- Less effects of the additional techniques
- Disadvantage: synthetic data generation is costly

## Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS") and eBay Inc. The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

## Links

Paper

Twitter