# Neural and Pre-trained MT

- Neural MT systems were already dominant even before pre-training came along
- Pre-training helps…but because systems like GPT-4 do less well on languages other than English, LLMs haven't revolutionized MT as much as other subtasks

# Transformer MT

| Model | BLEU | |
|---|---|---|
| | EN-DE | EN-FR |
| ByteNet [18] | 23.75 | |
| Deep-Att + PosUnk [39] | | 39.2 |
| GNMT + RL [38] | 24.6 | 39.92 |
| ConvS2S [9] | 25.16 | 40.46 |
| MoE [32] | 26.03 | 40.56 |
| Deep-Att + PosUnk Ensemble [39] | | 40.4 |
| GNMT + RL Ensemble [38] | 26.30 | 41.16 |
| ConvS2S Ensemble [9] | 26.36 | **41.29** |
| Transformer (base model) | 27.3 | 38.1 |
| Transformer (big) | **28.4** | **41.8** |

- Big = 6 layers, 1000 dim for each token, 16 heads, base = 6 layers + other params halved
- GNMT: Large LSTM system with attention; even the first version of Transformers already beat this!

Vaswani et al. (2017)

# Frontiers in MT: Small Data

| ID | system | BLEU 100k | BLEU 3.2M |
|---|---|---|---|
| 1 | phrase-based SMT | 15.87 ± 0.19 | 26.60 ± 0.00 |
| 2 | NMT baseline | 0.00 ± 0.00 | 25.70 ± 0.33 |
| 3 | 2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing) | 7.20 ± 0.62 | 31.93 ± 0.05 |
| 4 | 3 + reduce BPE vocabulary (14k → 2k symbols) | 12.10 ± 0.16 | - |
| 5 | 4 + reduce batch size (4k → 1k tokens) | 12.40 ± 0.08 | 31.97 ± 0.26 |
| 6 | 5 + lexical model | 13.03 ± 0.49 | 31.80 ± 0.22 |
| 7 | 5 + aggressive (word) dropout | 15.87 ± 0.09 | **33.60** ± 0.14 |
| 8 | 7 + other hyperparameter tuning (learning rate, model depth, label smoothing rate) | **16.57** ± 0.26 | 32.80 ± 0.08 |
| 9 | 8 + lexical model | 16.10 ± 0.29 | 33.30 ± 0.08 |

- Synthetic small data setting: German -> English
- Even with 100,000 examples, a well-tuned neural system can do on par with phrase-based models

Sennrich and Zhang (2019)

# Frontiers in MT: Low-Resource

- Lots of interest in deploying MT systems for languages with little or no parallel data

- BPE allows us to transfer models even without training on a specific language

- Pre-trained models can help further

Burmese, Indonesian, Turkish

| | BLEU | | |
|---|---|---|---|
| Transfer | My→En | Id→En | Tr→En |
| baseline (no transfer) | 4.0 | 20.6 | 19.0 |
| transfer, train | 17.8 | 27.4 | 20.3 |
| transfer, train, reset emb, train | 13.3 | 25.0 | 20.0 |
| transfer, train, reset inner, train | 3.6 | 18.0 | 19.1 |

Table 3: Investigating the model's capability to restore its quality if we reset the parameters. We use En→De as the parent.

Aji et al. (2020)

# Frontiers in MT: Low-Resource

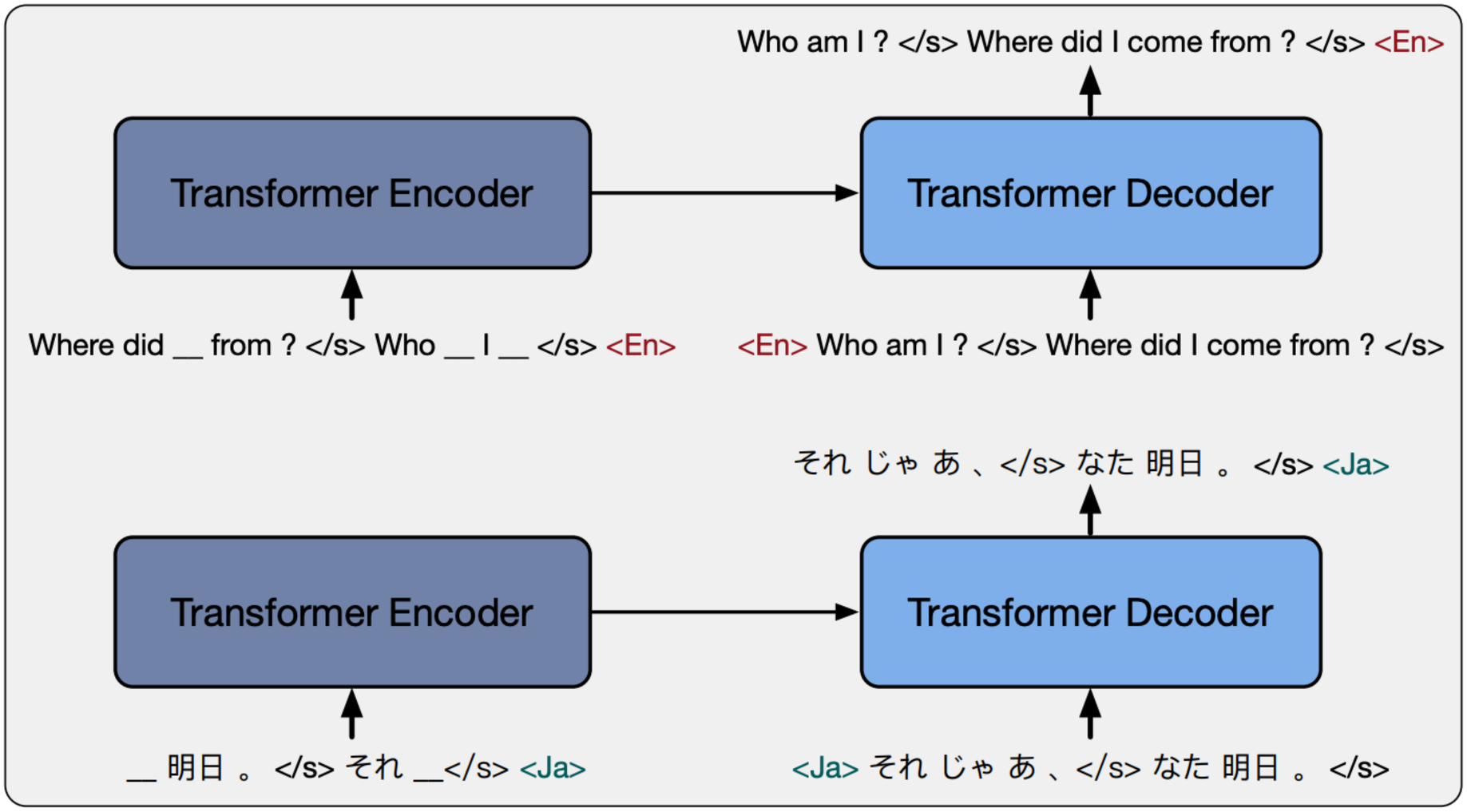
| Transferring | | BLEU | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | De→En parent | | | En→De parent | | | |
| Emb. | Inner | My→En | Id→En | Tr→En | My→En | Id→En | Tr→En | avg. |
| Y | Y | 17.8 | 27.4 | 20.3 | 17.5 | 27.5 | 20.2 | 21.7 |
| N | Y | 13.6 | 25.3 | 19.4 | 10.8 | 24.9 | 19.3 | 18.3 |
| Y | N | 3.0 | 18.2 | 19.1 | 3.4 | 18.8 | 18.9 | 13.7 |
| N | N | 4.0 | 20.6 | 19.0 | 4.0 | 20.6 | 19.0 | 14.5 |

Table 2: Transfer learning performance by only transferring parts of the network. Inner layers are the non-embedding layers. N = not-transferred. Y = transferred.

- Very important to transfer the basic Transformer “skills”, but re-learning the embeddings seems fine in many cases

Aji et al. (2020)

# Multilingual Models

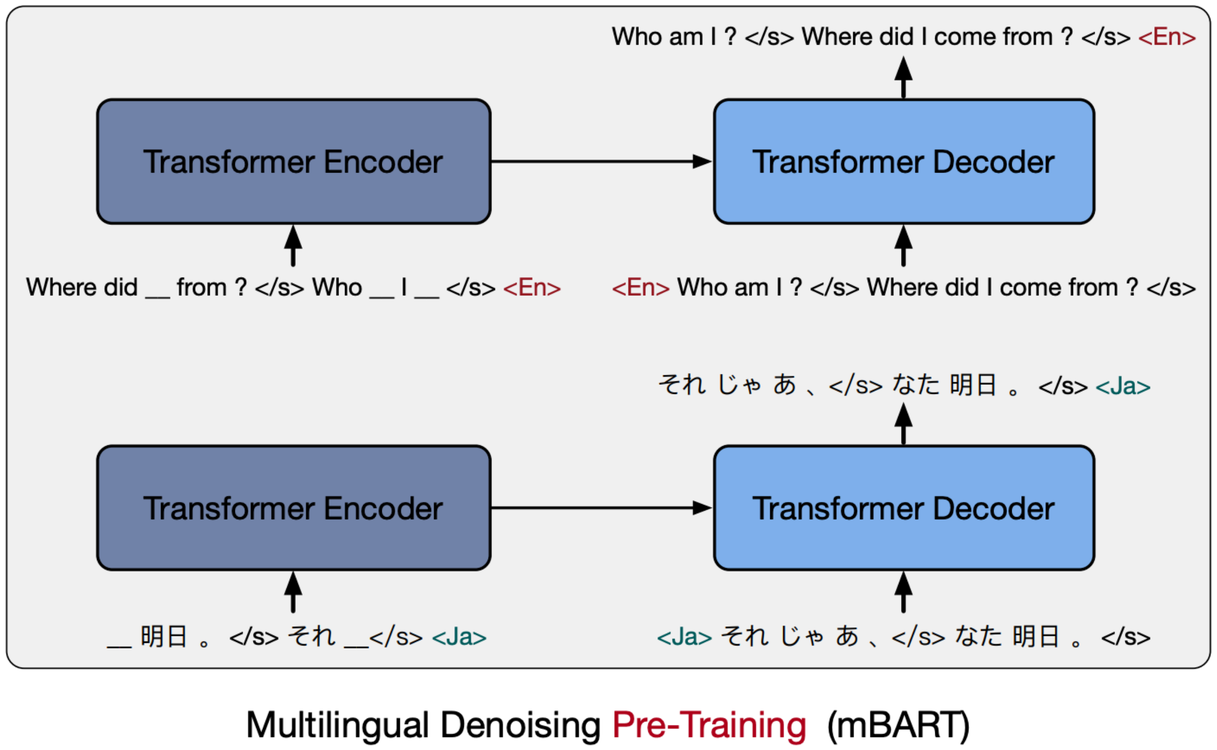



Multilingual Denoising Pre-Training (mBART)

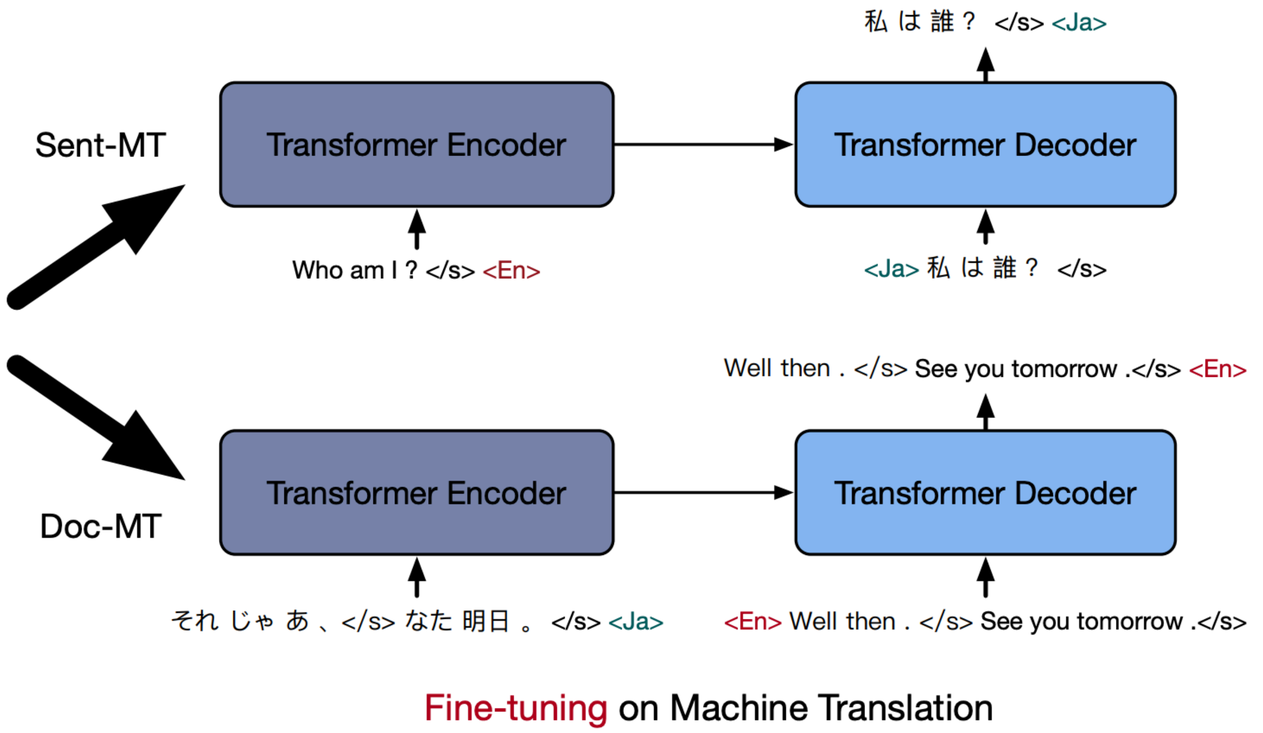
- mBART: pre-trained model using the BART objective, where inputs and outputs are augmented with language codes and many languages are learned in a single model

Yinhan Liu et al. (2020)

# Multilingual Models




Multilingual Denoising Pre-Training (mBART)




Fine-tuning on Machine Translation

Yinhan Liu et al. (2020)

# mBART Results

| Languages | En-Gu | | En-Kk | | En-Vi | | En-Tr | | En-Ja | | En-Ko | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Source | WMT19 | | WMT19 | | IWSLT15 | | WMT17 | | IWSLT17 | | IWSLT17 | |
| Size | 10K | | 91K | | 133K | | 207K | | 223K | | 230K | |
| Direction | ← | → | ← | → | ← | → | ← | → | ← | → | ← | → |
| Random | 0.0 | 0.0 | 0.8 | 0.2 | 23.6 | 24.8 | 12.2 | 9.5 | 10.4 | 12.3 | 15.3 | 16.3 |
| mBART25 | **0.3** | **0.1** | **7.4** | **2.5** | **36.1** | **35.4** | **22.5** | **17.8** | **19.1** | **19.4** | **24.6** | **22.6** |

| Languages | En-Nl | | En-Ar | | En-It | | En-My | | En-Ne | | En-Ro | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Source | IWSLT17 | | IWSLT17 | | IWSLT17 | | WAT19 | | FLoRes | | WMT16 | |
| Size | 237K | | 250K | | 250K | | 259K | | 564K | | 608K | |
| Direction | ← | → | ← | → | ← | → | ← | → | ← | → | ← | → |
| Random | 34.6 | 29.3 | 27.5 | 16.9 | 31.7 | 28.0 | 23.3 | 34.9 | 7.6 | 4.3 | 34.0 | 34.3 |
| mBART25 | **43.3** | **34.8** | **37.6** | **21.6** | **39.8** | **34.0** | **28.3** | **36.9** | **14.5** | **7.4** | **37.8** | **37.7** |

- Random = random initialization of Transformer
- mBART does much better, but still doesn't solve MT in low-resource languages (Gujarati, Kazakh)

Ouyang et al. (2022)

# mBART Results

| | |
|---|---|
| **SOURCE Zh** | 针对政府的沉默态度,初级医生委员会执行委员会已于今日正式要求英国医学协会理事会召开特别会议批准旨在从九月初开始升级劳工行动的一项长期计划。 |
| **TARGET En** | In response to the government's silence, JDC exec has today made a formal request for a special meeting of BMA Council to authorise a rolling programme of escalated industrial action beginning in early September. |
| **mBART25 Ja-En** | In response to the government's silence, the Council of Chief Medical Officers has formally requested today the Royal College of Physicians to hold a special meeting to approve a long-term workforce action that starts in September. |
| **mBART25 Ko-En** | In response to the government's silence, the Chief Medical Officers' Council is calling today for a special session at the Council of the British Medical Association, which is a long-term initiative to upgrade labor from September. |
| **mBART25 Zh-En** | In response to the government's silence, the Board of Primary Doctors has today formally asked the British Medical Association to hold a special meeting to approve a long-term plan that starts in the beginning of September. |

Ouyang et al. (2022)

# Evaluating Translations with LLMs

```
Score the following translation from {source_lang} to {target_lang} with respect
to the human reference on a continuous scale from 0 to 100, where score of zero means
"no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source_lang} source: "{source_seg}"
{target_lang} human reference: {reference_seg}
{target_lang} translation: "{target_seg}"
Score:
```

Figure 1: The best-performing prompt based on Direct Assessment expecting a score between 0–100. Template **portions in bold face** are used only when a human reference translation is available.

- Outperforms many learned MT metrics, like Transformers trained over (source, target, reference) triples to reproduce human judgments

Kocmi et al. (2023)