# Multilingual Pre-Training and Cross-Lingual Transfer for MT and NLP

## Erweiterungsmodul: Machine Translation
## Sommersemester 2023

Katharina Hämmerl
haemmerl@cis.lmu.de

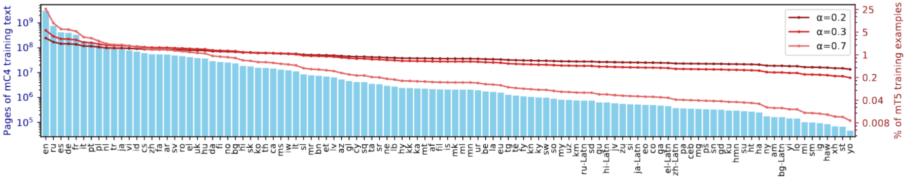
LMU München, Center for Information and Language Processing

August 02, 2023

# Motivation

# Data matters

x10000




x125

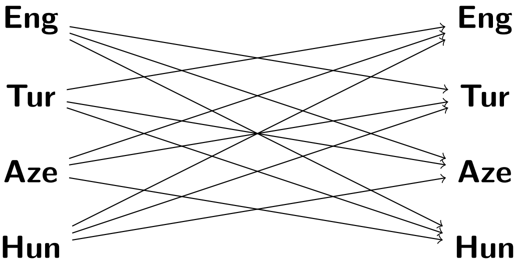
Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents $\alpha$ (right axis). Our final model uses $\alpha$=0.3.

- mC4 dataset, from mT5 paper
- Monolingual datasets → Situation is at least this bad for parallel data

Xue et al. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL 2021

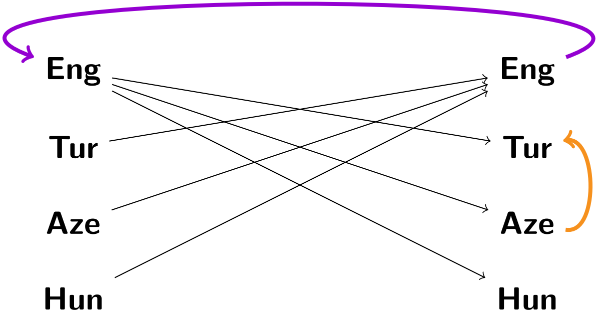
*First part of slides adapted from Xinyi Wang, CMU*

# Supporting many language pairs is hard




→ Just translating from 4 to 4 languages requires 4*3=12 NMT models

# Supporting many language pairs is hard

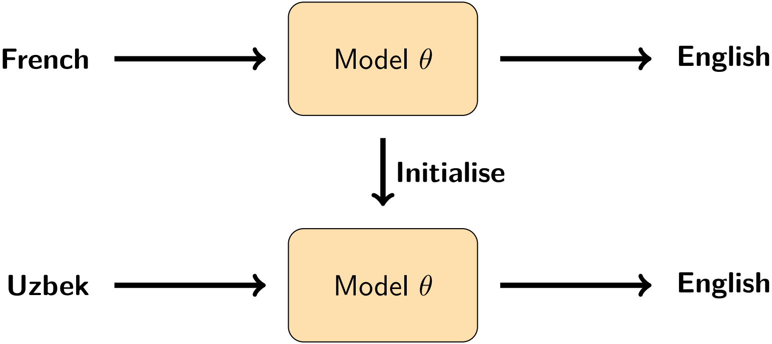



→ Instead: pivot translation, but this adds time and can introduce extra errors

→ Related but low-resource language pairs suffer especially

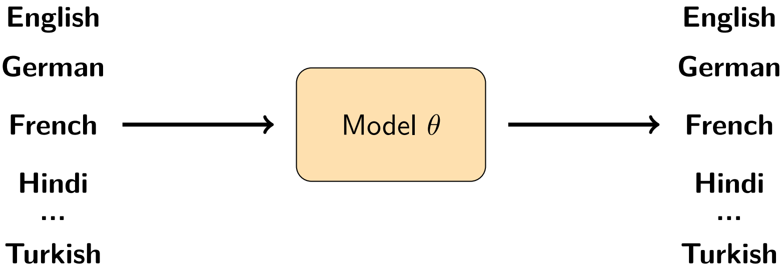
# Towards Multilingual MT

# Cross-Lingual Transfer




- Train a model on high-resource language pair
- Finetune on small low-resource language pair

Zoph et al. 2016. Transfer learning for low-resource neural machine translation. EMNLP 2016.




- Train a single model on a mixed dataset from multiple languages (e.g., five languages in the paper)

Johnson et al. 2017. Google's multilingual neural machine translation system: Enabling Zero-Shot Translation. TACL.

# Multilingual Training

**<2fr> How are you?** → **Comment ça va?**

**<2tr> How are you?** ⟶ Model $\theta$ ⟶ **Nasılsın?**

**...** **...**

**<2es> How are you?** → **Cómo estás?**

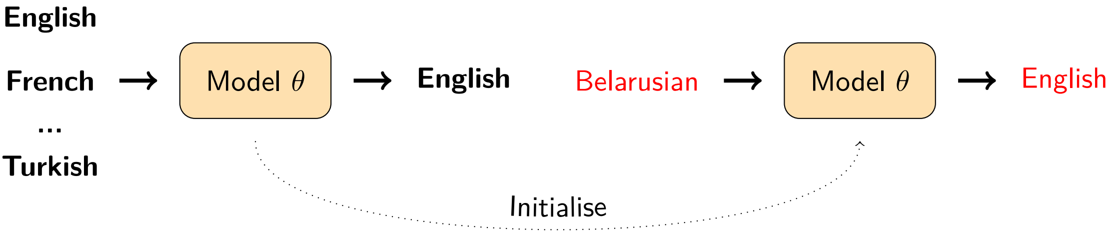
- NMT needs to generate into many languages, simply add target language label

Johnson et al. 2017. Google's multilingual neural machine translation system: Enabling Zero-Shot Translation. TACL.

# Combining the Two Methods

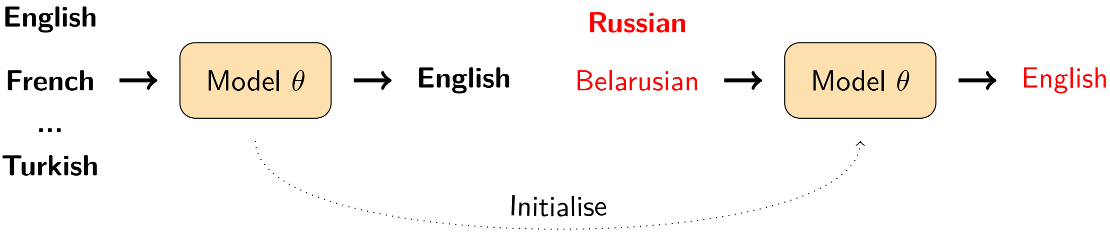
- We just covered the two main paradigms for multilingual methods
  - Cross-lingual transfer
  - Multilingual training
- How best to combine the two to train a good model for a new language?

# Rapid Adaptation to New Languages




- First, do multilingual training on many languages (eg. 58 languages in the paper)
- Next fine-tune the model on a new low-resource language

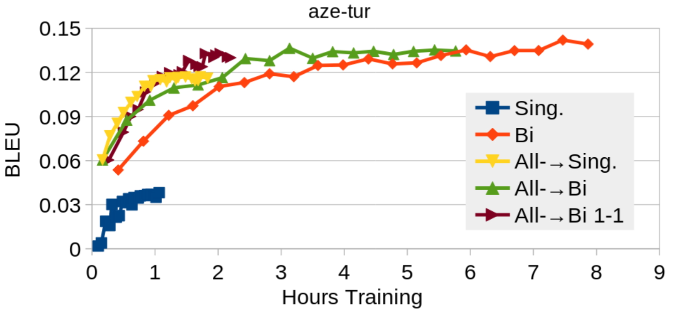
Neubig and Hu. 2018. Rapid adaptation of Neural Machine Translation to New Languages. EMNLP 2018.




- Regularized fine-tuning: fine-tune on low-resource language and its related high-resource language to avoid overfitting

Neubig and Hu. 2018. Rapid adaptation of Neural Machine Translation to New Languages. EMNLP 2018.

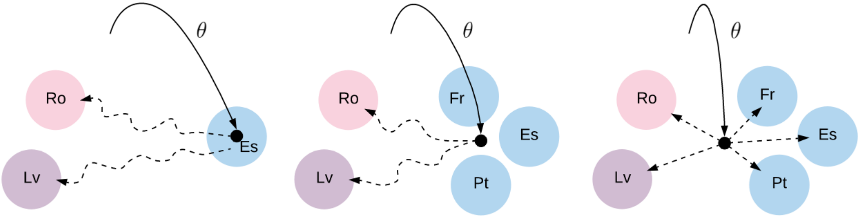
# Rapid Adaptation to New Languages




- All → xx models: adapting from a multilingual model makes convergence faster
- Regularized fine-tuning yields better final performance

Neubig and Hu. 2018. Rapid adaptation of Neural Machine Translation to New Languages. EMNLP 2018.

# Meta-Learning for Multilingual Training




(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

- Learning a good initialization of model for fast adaptation to all languages
- Inner loop: optimize/learn for each language
- Outer loop (meta objective): learn how to quickly optimize for each language

Gu et al. 2018. Meta-learning for low-resource neural machine translation. EMNLP 2018.

## Zero-shot Transfer

- Train models that work for a language without annotated data in that language
- Allowed to train using **monolingual** data for the test language or **annotated data for other languages**

# Zero-shot Transfer in MT

**Zulu - English** ⟵ some Bible data

**Italian - English** ⟵ News, European Parliament documents,....

**Zulu - Italian** ⟵ not much data available

→ Parallel data are English centric

# Zero-shot Transfer in MT

**Training:**

<2en> Zulu-English src → Model $\theta$ → Zulu-English tgt

<2en> Italian-English src → Model $\theta$ → Italian-English tgt

<2it> English-Italian src → Model $\theta$ → English-Italian tgt

**Testing:**

<2it> Sawubona → Model $\theta$ → Ciao

# Zero-Shot Transfer in MT

- Multilingual training allows zero-shot transfer
- Train on {Zulu-English, English-Zulu, English-Italian, Italian-English}
- Zero-shot: Translate Zulu to Italian without Zulu-Italian parallel data

Johnson et al. 2017. Google's multilingual neural machine translation system: Enabling Zero-Shot Translation. TACL.

# Improving Zero-Shot Transfer in NMT: Noised Monolingual Data

**Training:**

<2en> Zulu-English src → Model $\theta$ → Zulu-English tgt

<2it> **Noised Italian** → Model $\theta$ → **Italian**

**Testing:**

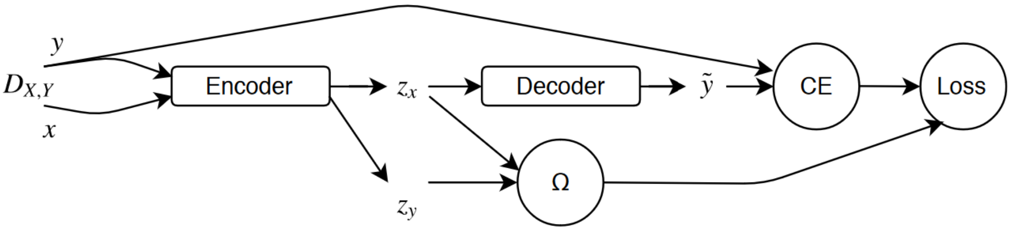
<2it> Sawubona → Model $\theta$ → Ciao

- Add monolingual data by asking the model to reconstruct the noisy version of the monolingual data
- Use masked language model objective

Siddhant et al. 2020. Leveraging Monolingual Data with Self-Supervision for Multilingual NMT. ACL 2020.

Tang et al. 2021. Multilingual Translation from Denoising Pre-Training. ACL Findings 2021.

# Improving Zero-Shot Transfer in NMT: Alignment of Multilingual Representations




- Translation objective alone might not encourage language-invariant representation
- Add an extra loss to align source and target encoder representation

Arivazhagan et al. 2019. The Missing Ingredient in Zero-Shot Neural Machine Translation. arXiv, CoRR.

# Multilingual Pre-Trained Models

# Multilingual Pre-Training

- We've been talking about multilingual MT specifically
- Pre-training (on monolingual data) is used in MT to get better language modelling, better results
- Pre-training is a generalisable principle
- Multilingual, monolingual, encoder, decoder,...

→ Kind of a detour from MT, but we'll come back around!

# Why Multilingual Pre-Training?

- Reusable models for multiple languages
- Fewer resources than maintaining individual models
- Faster adaptation or no adaptation to use for different languages
- Better for lower-resource languages than training individual models
- Can model languages where there is not enough data for a monolingual model

# Recall: Encoder, Decoder, Encoder-Decoder

## Encoder-Only

- Typically trained on masked language modelling or similar
- Outputs vectors/matrices
- Fine-tuned for, e.g., classification tasks
- Includes BERT-type models

## Encoder-Decoder

- Trained on sequence-to-sequence data, or e.g. span corruption
- Outputs text
- Can be fine-tuned for various tasks
- Includes (most) MT models

## Decoder-Only

- Typically trained on autoregressive LM or similar
- Outputs text
- Often used with prompts and in-context learning
- Includes GPT-type models

# mBERT and XLM-R

- Two similar, famous **encoder** models
- mBERT supports 104 languages, XLM-R 100.
- Both: Concatenate data from all training languages → MLM
- XLM-R is trained on more data, better optimised, has a Large version (more recently, up to XXL)
- Show cross-lingual representations despite **no explicit** cross-lingual signal
- Due to overlapping tokens, compression/limited capacity,...?

Devlin et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.
Conneau et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020.
Dufter and Schütze. 2020. Identifying Elements Essential for BERT's Multilinguality. EMNLP 2020.
Goyal et al. Larger-Scale Transformers for Multilingual Masked Language Modeling. Repl4NLP 2021.

# Zero-Shot Cross-Lingual Transfer

**Pre-Training:**

**English** … **Malay** $\longrightarrow$ Model $\theta$ $\longrightarrow$ **Encoder Representation**

**Fine-Tuning:**

**English** sentence $\longrightarrow$ Model $\theta$ + Head $\longrightarrow$ **POS tags**

**Testing:**

**Malay** sentence $\longrightarrow$ Model $\theta$ + Head $\longrightarrow$ **POS tags**

# Transfer Performance

| Task | Model | EN | ZH Δ | TR Δ | RU Δ | AR Δ | HI Δ | EU Δ | FI Δ | HE Δ | IT Δ | JA Δ | KO Δ | SV Δ | VI Δ | TH Δ | ES Δ | EL Δ | DE Δ | FR Δ | BG Δ | SW Δ | UR Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DEP | B | 91.2 | -43.9 | -46.0 | -28.1 | -56.4 | -36.1 | -50.2 | -30.7 | -36.1 | -17.1 | **-60.1** | -56.1 | -14.3 | - | - | - | - | - | - | - | - | - |
| | X | 92.0 | **-85.4** | -44.2 | -29.7 | -54.6 | -39 | -49.5 | -26.7 | -39 | -23.5 | -80.5 | -56.0 | -16.3 | - | - | - | - | - | - | - | - | - |
| POS | B | 95.8 | -38.0 | -35.9 | -16.0 | -40.1 | -33.4 | -34.6 | -21.9 | -33.4 | -19.8 | **-46.1** | -42.0 | -9.6 | - | - | - | - | - | - | - | - | - |
| | X | 96.3 | -69.2 | -27.7 | -14.3 | -37.1 | -27.3 | -31.9 | -17.9 | -27.3 | -19.0 | **-77.0** | -37.3 | -10.7 | - | - | - | - | - | - | - | - | - |
| NER | B | 92.4 | -23.3 | -11.6 | -10.7 | **-31.7** | -11.1 | -12.8 | -3.8 | -11.1 | -2.6 | -25.7 | -13.8 | -6.7 | - | - | - | - | - | - | - | - | - |
| | X | 91.6 | **-34.8** | -6.2 | -13.7 | -24.6 | -16.5 | -8.0 | -0.9 | -16.5 | -2.4 | -30.1 | -15.6 | -2.2 | - | - | - | - | - | - | - | - | - |
| XNLI | B | 82.8 | -13.6 | -20.6 | -13.5 | -17.3 | -21.3 | - | - | - | - | - | - | - | -11.9 | -28.1 | -8.1 | -14.1 | -10.5 | -7.8 | -13.3 | **-33.0** | -23.4 |
| | X | 84.3 | -11.0 | -11.3 | -9.0 | -13.0 | -14.2 | - | - | - | - | - | - | - | -9.7 | -12.3 | -5.8 | -8.9 | -7.8 | -6.1 | -6.6 | **-20.2** | -17.3 |
| XQuAD | B | 71.1 | -22.9 | -34.2 | -19.2 | -24.7 | -28.6 | - | - | - | - | - | - | - | -22.1 | **-43.2** | -16.6 | -28.2 | -14.8 | - | - | - | - |
| | X | 72.5 | **-26.2** | -18.7 | -15.4 | -24.1 | -22.8 | - | - | - | - | - | - | - | -19.7 | -14.8 | -14.5 | -15.7 | -16.2 | - | - | - | - |

Table 1: Zero-shot cross-lingual transfer performance on five tasks (DEP, POS, NER, XNLI, and XQuAD) with mBERT (B) and XLM-R (X). We show the monolingual EN performance and report drops in performance relative to EN for all target languages. Numbers in bold indicate the largest zero-shot performance drops for each task.

Lauscher et al. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. EMNLP 2020.

# How Language-Neutral Are These Models?

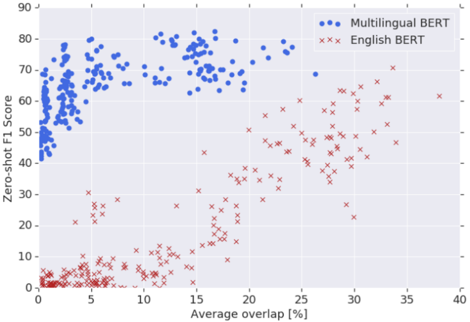
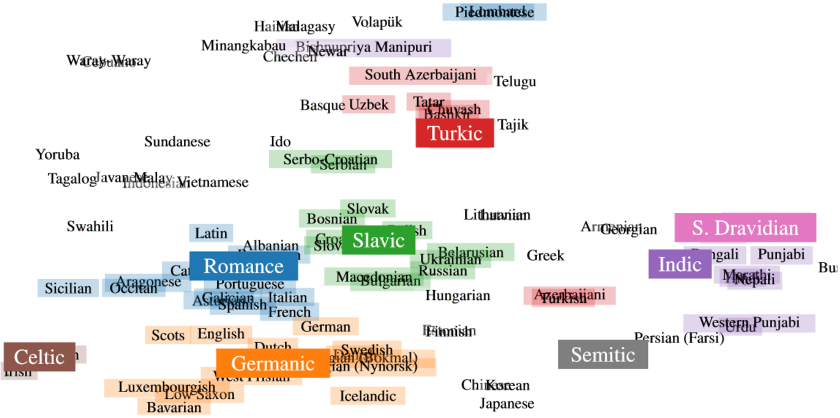



Figure 1: Zero-shot NER F1 score versus entity word piece overlap among 16 languages. While performance using EN-BERT depends directly on word piece overlap, M-BERT's performance is largely independent of overlap, indicating that it learns multilingual representations deeper than simple vocabulary memorization.

- x-axis: Average token overlap of the sequences with English
- Interpretation: Cross-lingual representation is responsible for better transfer performance in mBERT
- Works well even with different scripts for some pairs (Hindi-Urdu) but not others (English-Japanese)


Pires et al. 2019. How Multilingual is Multilingual BERT? ACL 2019.

# How Language-Neutral Are These Models?




Libovický et al. 2020. On the Language Neutrality of Pre-trained Multilingual Representations. EMNLP 2020.

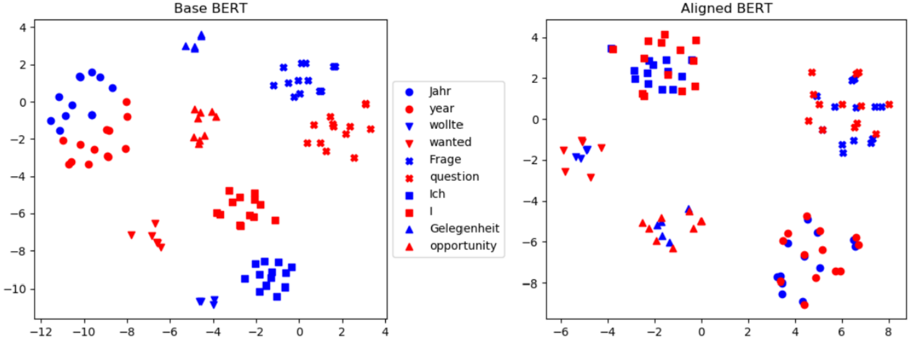
# How Language-Neutral Are These Models?

| Task | Model | SYN P | SYN S | PHON P | PHON S | INV P | INV S | FAM P | FAM S | GEO P | GEO S | SIZE P | SIZE S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DEP | XLM-R | 0.77 | 0.78 | **0.83** | **0.77** | 0.46 | -0.04 | 0.68 | 0.61 | 0.80 | 0.81 | 0.62 | 0.47 |
| | mBERT | **0.92** | **0.91** | 0.79 | 0.74 | 0.55 | -0.01 | 0.76 | 0.62 | 0.64 | 0.69 | 0.79 | 0.59 |
| POS | XLM-R | 0.68 | 0.79 | **0.81** | **0.81** | 0.38 | 0.02 | 0.58 | 0.74 | 0.80 | 0.73 | 0.54 | 0.46 |
| | mBERT | **0.90** | **0.87** | 0.86 | 0.81 | 0.57 | 0.02 | 0.82 | 0.80 | 0.66 | 0.72 | 0.47 | 0.39 |
| NER | XLM-R | 0.49 | 0.49 | **0.80** | **0.83** | 0.27 | 0.14 | 0.47 | 0.55 | 0.77 | 0.81 | 0.37 | 0.35 |
| | mBERT | 0.60 | 0.74 | **0.81** | **0.84** | 0.34 | -0.04 | 0.53 | 0.58 | 0.59 | 0.73 | 0.42 | 0.38 |
| XNLI | XLM-R | **0.88** | **0.90** | 0.29 | 0.27 | 0.31 | -0.11 | 0.63 | 0.54 | 0.54 | 0.74 | 0.70 | 0.76 |
| | mBERT | **0.87** | 0.86 | 0.21 | 0.08 | 0.29 | 0.04 | 0.61 | 0.47 | 0.55 | 0.67 | 0.77 | **0.91** |
| XQuAD | XLM-R | 0.69 | 0.53 | **0.85** | **0.81** | 0.62 | -0.01 | **0.81** | 0.54 | 0.43 | 0.50 | **0.81** | 0.55 |
| | mBERT | 0.84 | 0.89 | 0.56 | 0.48 | 0.55 | 0.22 | 0.79 | 0.64 | 0.51 | 0.55 | **0.89** | **0.96** |

Table 2: Correlations between zero-shot transfer performance with mBERT and XLM-R for different downstream tasks with linguistic proximity features (SYN, PHON, INV, FAM and GEO) and pretraining size of target-language corpora (SIZE). Results reported in terms of Pearson (P) and Spearman (S) correlation coefficients.

Lauscher et al. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. EMNLP 2020.

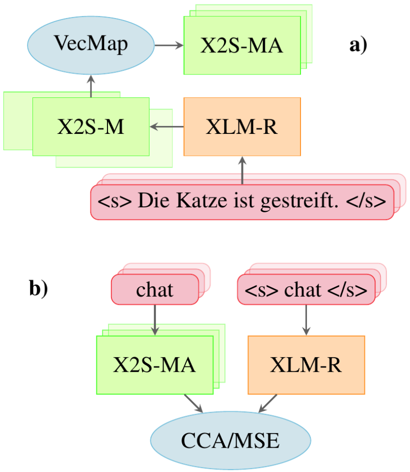
# Aligning Representations in Multilingual Models




- Minimise distance between aligned words in parallel text
- Regularise to stay close to initial representations

Cao et al. 2020. Multilingual Alignment of Contextual Word Representations. ICLR 2020.

# Aligning Representations in Multilingual Models




- Part of the model's appeal is training without parallel data. How can we align without resorting to parallel text?
- Extracted static embeddings from the model and applied traditional embedding alignment
- Minimise distance between contextual word embeddings and aligned static embeddings
- Regularise by adding masked language modelling

Hämmerl et al. 2022. Combining Static and Contextualised Multilingual Embeddings. ACL Findings 2022

# The "Curse of Multilinguality"

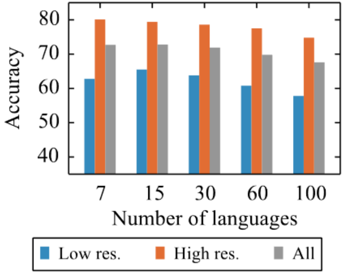



Figure 2: The transfer-interference trade-off: Low-resource languages benefit from scaling to more languages, until dilution (interference) kicks in and degrades overall performance.

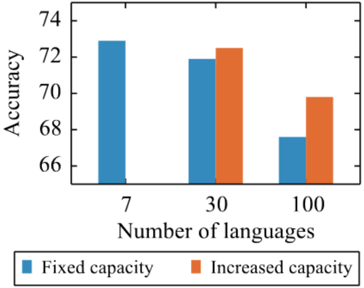
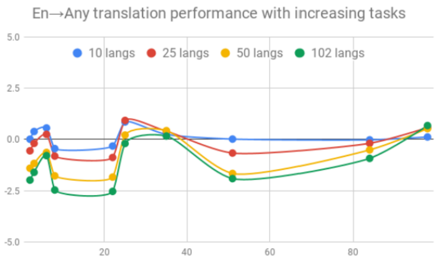
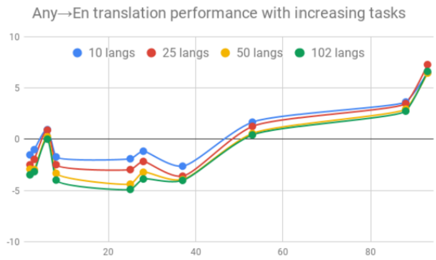



Figure 4: Adding more capacity to the model alleviates the curse of multilinguality, but remains an issue for models of moderate size.

Conneau et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020.

# ..and what it looks like in MT

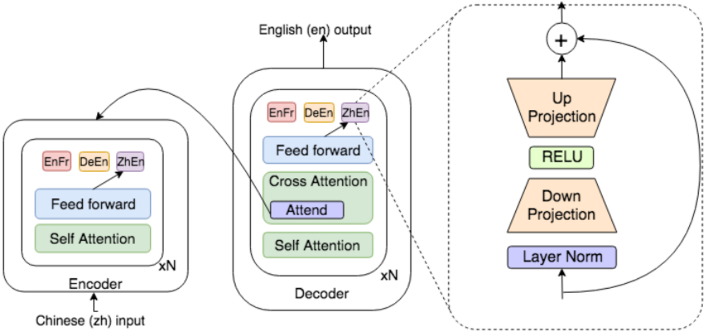



- x-axis: Rank of language w.r.t. data size—10 languages plotted
- y-axis: BLEU score relative to bilingual models
- Interpretation: Lower-resource languages benefit more from multilingual training, high-resource languages suffer. All get worse as language pairs added

Arivazhagan et al. 2019. Massively Multilingual Neural Machine Translation in the Wild. arXiv, CoRR.

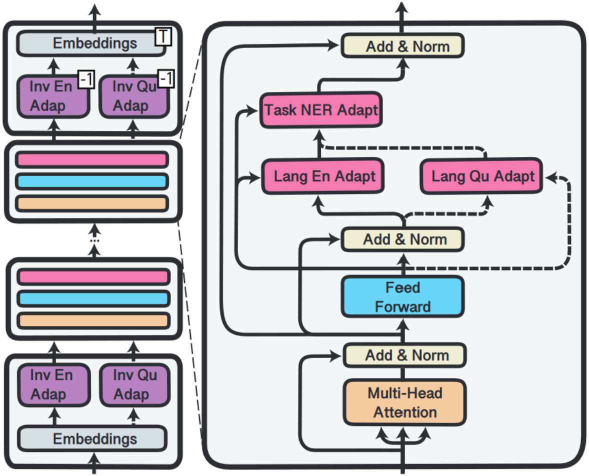
# One Solution: Adding Language-Specific Layers




- Add a small module for each language pair ($\sim$ adapter concept)
- Much better at matching bilingual baseline for high-resource languages

Bapna et al. 2019. Simple, Scalable adaptation for neural machine translation. EMNLP 2019.

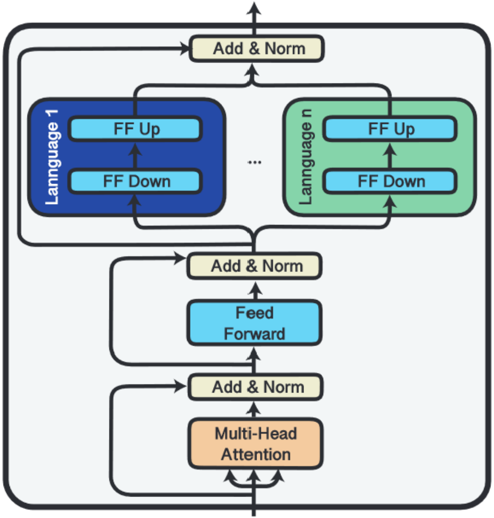
# MAD-X Framework




- Cross-lingual transfer by training task adapters and language adapters & combining them
- Swap language adapters for transfer
- Plus invertible adapters for embeddings

Pfeiffer et al. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. EMNLP 2020.

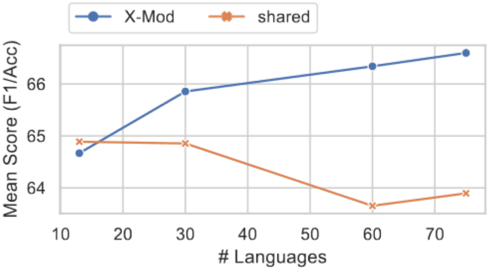
# Modular Transformers




- Schematic of a modular transformer layer (green and dark blue are language modules)
- Allocate some fraction of parameters to each language
- Unlike previous slides, already pre-train with modules
- Can add further languages by training more modules


Pfeiffer et al. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. NAACL 2022.

# Modular Transformers




(b) Mean Performance on XNLI and NER.

- Left: NER and XNLI performance
- Why is this unfair?
- SHARED model has one module for all languages, but the more languages, the more modules for X-MOD
- Still promising!
- Disadvantage: Swapping parameters, need to always know language

Pfeiffer et al. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. NAACL 2022.

# Multilingual vs Cross-Lingual Pre-Training

- There are other models that do make use of parallel data!
- Terminology, roughly: *cross-lingual training* = parallel data; *multilingual training* = not necessarily parallel
- Different possible objectives, for example:
  - Translation Language Modelling (TLM) concatenates translation pair in input → MLM. Also others that are versions of monolingual objectives

    Lample and Coneau. 2019. Cross-lingual Language Model Pretraining. NeurIPS 2019.
  - Cross-Lingual Contrastive Learning (XLCO): Maximise sequence-level mutual information between parallel sentences

    Chi et al. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. NAACL 2021.

# Recall: Encoder, Decoder, Encoder-Decoder

## Encoder-Only

- Typically trained on masked language modelling or similar
- Outputs vectors/matrices
- Fine-tuned for, e.g., classification tasks
- Includes BERT-type models

## Encoder-Decoder

- Trained on sequence-to-sequence data, or e.g. span corruption
- Outputs text
- Can be fine-tuned for various tasks
- Includes (most) MT models

## Decoder-Only

- Typically trained on autoregressive LM or similar
- Outputs text
- Often used with prompts and in-context learning
- Includes GPT-type models

# Multilingual Encoder-Decoder Models (Examples)

- mBART: Denoising Auto-Encoder for 25/50 langs. Can be fine-tuned for MT

  Liu et al. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. TACL.

- mT5: Span masking, models in different sizes, for 101 languages (C4 corpus). They show fine-tuned results for XTREME benchmark

  Xue et al. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. NAACL 2021.

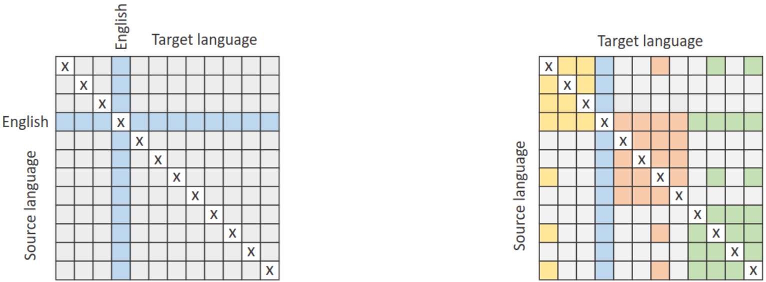
- nmT5: Similar setup, but add parallel data with (denoised) NMT objective

  Kale et al. 2021. nmT5 - Is parallel data still relevant for pre-training massively multilingual language models? ACL 2021.

- M2M-100: Many-to-many parallel data training for 100 languages

  Fan et al. 2021. Beyond English-Centric Multilingual Machine Translation. JMLR 2021.

# M2M-100: More Details

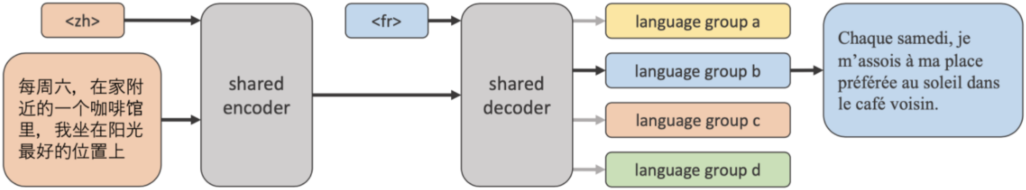



(a) English-Centric Multilingual

(b) M2M-100: Many-to-Many Multilingual Model

- Used mined many-to-many data partly from existing corpora, partly extended themselves
- Worked with language groupings to constrain global search, as well as *bridge languages* between groups

# M2M-100: More Details



Translating from Chinese to French with Dense + Language-Specific Sparse Model

- Also add language-specific ("sparse") layers to the model
- Group languages with less than 100M sentences

→ Increase capacity but training/inference time stays similar

→ Largest model they train this way has 15.4B parameters

# Multilinguality in LLMs

# Recall: Encoder, Decoder, Encoder-Decoder

## Encoder-Only

- Typically trained on masked language modelling or similar
- Outputs vectors/matrices
- Fine-tuned for, e.g., classification tasks
- Includes BERT-type models

## Encoder-Decoder

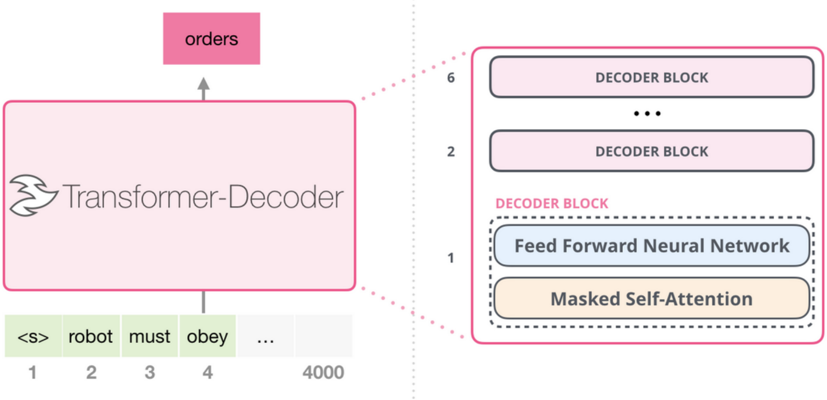
- Trained on sequence-to-sequence data, or e.g. span corruption
- Outputs text
- Can be fine-tuned for various tasks
- Includes (most) MT models

## Decoder-Only

- Typically trained on autoregressive LM or similar
- Outputs text
- Often used with prompts and in-context learning
- Includes GPT-type models

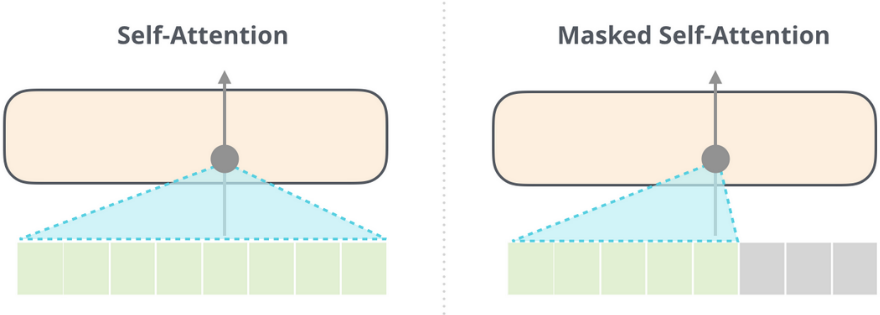
# Pre-Trained Decoder Models

- Many "Large Language Models" (LLMs) are *decoder-only*
- How to train a decoder-only model?




Jay Alammar. 2018. The Illustrated GPT-2. Blog post.

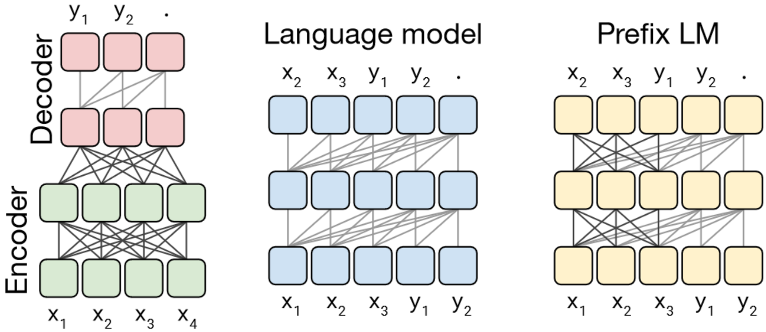
# Pre-Trained Decoder Models




- *Autoregressive* (start-to-end of sequence) training objective
- Unlike BERT, does not have bi-directional context; context after current token is completely masked out for self-attention
- Difference in training between GPT-2 and later versions is in details

Jay Alammar. 2018. The Illustrated GPT-2. Blog post.

# Side Note: Causal LM vs Prefix LM




- Alternative to masking options we already know: Prefix LM
- Full attention over an input sequence (similar to encoder); left-to-right over target

Raffel et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2020.

# Proprietary Models

- LLMs today take huge amounts of resources
- So training is most often done by or with huge companies/organisations
- OpenAI (GPT-3/4, ChatGPT), Google (PaLM 1/2, Bard),.. have trained *closed* models
- There are technical reports that reveal *some* information and advertise evaluation results
- But they are *not publicly released* and *not reproducible*
- Even running inference would be a challenge (!)

# Open-Source Model Requirements

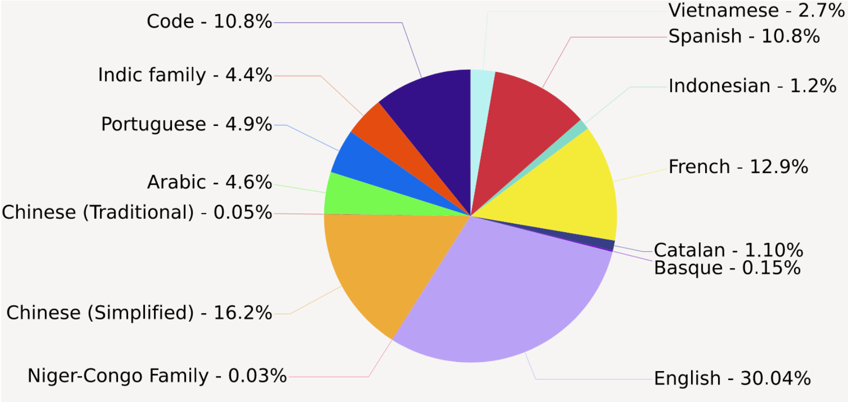
What does an open model look like?

- A way to access, reproduce, search the training data
- Should also be *documented* for understanding without going through all of it
- Detailed training documentation: number of parameters, hyperparameters, resources,..
- A way to download, re-train, inspect the model weights
- Ideally *also* a demo/API running somewhere
- Appropriate licensing

→ Many models have some but not all of these

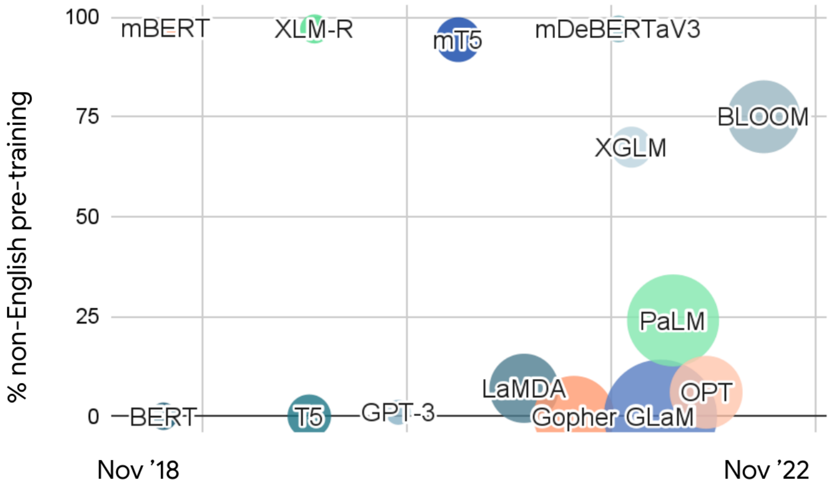
! We cannot do science without model transparency

# BigScience BLOOM




BigScience Workshop. 2022. BLOOM Model Card. BigScience Workshop. Web page.

# State of Multilingual LLMs




Sebastian Ruder. 2022. The State of Multilingual AI. Blog post.

# Documenting Data

- Models are only as good as their data
- We need to know who is represented, what kind of language is in there (varieties, toxicity, biases,..)
- We need to know if test sets are in the pre-training data (contamination)
- And more!

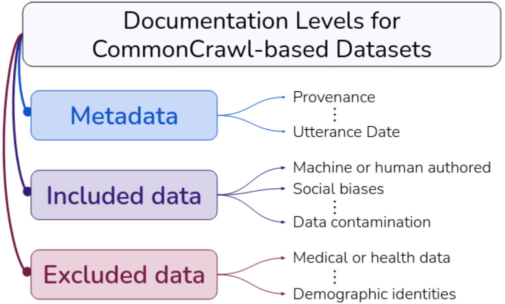



Figure 1: We advocate for three levels of documentation when creating web-crawled corpora. On the right, we include some example of types of documentation that we provide for the C4.EN dataset.

Dodge et al. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. EMNLP 2021

- A 2022 audit of multiple multilingual corpora found significant problems in quality of low-resource language data in particular
- Had speakers of 70 languages rate 100 lines per audited sub-corpus (sometimes based on educated guesses)
- Labelled “correct” data vs. multiple categories of issues

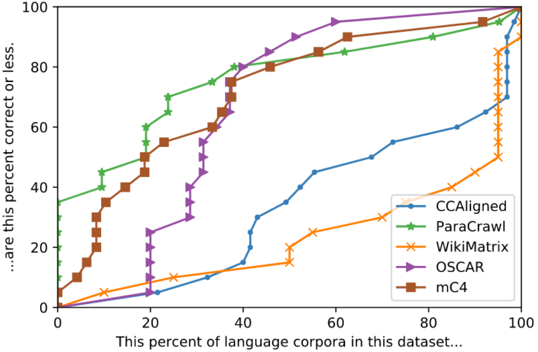



Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

Kreutzer et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. TACL.

# How Well Can LLMs Translate?

- Multiple studies collecting data points/snapshots of MT quality in LLMs
- Typically look at 0-shot, 1-shot, 5-shot
- 5-shot does reasonably well
- GPT-3.5 showed good/competitive results on a few very high-resource pairs
- But it did poorly on low-resource pairs and a direct-translation pair compared to WMT models

Hendy et al. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv, CoRR.

- Good results from PaLM-540B in another paper, but only evaluated on few high-resource pairs
- Not quite competitive with the WMT systems chosen in this paper

Vilar et al. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. ACL 2023

# Summary

# Summary

- Introduced cross-lingual transfer from an MT perspective
- Discussed multilingual training and adaptation
- Zero-shot transfer
- Expanded to pre-trained multilingual models more generally
- Discussed language neutrality and transfer performance
- Adapters and modular models
- Situated multilingual LLMs
- Highlighted issues around data and documentation

## Summary

- Very broad view of multilingual pre-training, cross-lingual transfer, and related topics
- Aim to give a high-level view/understanding
- Many papers mentioned
  $\rightarrow$ Possible starting points for learning more
- Thanks for listening even after end of term!