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
1 Motivation

2 Towards Multilingual MT

3 Multilingual Pre-Trained Models

4 Multilinguality in LLMs

5 Summary
Motivation

1. Motivation
2. Towards Multilingual MT
3. Multilingual Pre-Trained Models
4. Multilinguality in LLMs
5. Summary
Data matters

- mC4 dataset, from mT5 paper
- Monolingual datasets \(\rightarrow\) 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 $4 \times 3 = 12$ 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
1 Motivation

2 Towards Multilingual MT

3 Multilingual Pre-Trained Models

4 Multilinguality in LLMs

5 Summary
Cross-Lingual Transfer

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

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

Multilingual Training

- `<2fr>` How are you?
- `<2tr>` How are you?
- `<2es>` How are you?

Comment ça va?
Naslın?
Cómo estás?

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

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?
First, do multilingual training on many languages (eg. 58 languages in the paper)

Next fine-tune the model on a new low-resource language

Rapid Adaptation to New Languages

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

Rapid Adaptation to New Languages

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

- 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

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  \[\text{Model } \theta\]  \rightarrow  Zulu-English tgt
<2en> Italian-English src  \rightarrow  Italian-English tgt
<2it> English-Italian src  \rightarrow  English-Italian tgt

Testing:

<2it> Sawubona  \rightarrow  Model } \theta\]  \rightarrow  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

Improving Zero-Shot Transfer in NMT: Noised Monolingual Data

**Training:**

<2en> Zulu-English src  Zulu-English tgt

<2it> Noised Italian  Model  θ  Italian

**Testing:**

<2it> Sawubona  Model  θ  Ciao

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

---


Translation objective alone might not encourage language-invariant representation

Add an extra loss to align source and target encoder representation

Multilingual Pre-Trained Models

1. Motivation
2. Towards Multilingual MT
3. Multilingual Pre-Trained Models
4. Multilinguality in LLMs
5. Summary
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 $\rightarrow$ 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,...?

Zero-Shot Cross-Lingual Transfer

Pre-Training:
- English

... → Model $\theta$ → Encoder Representation

- Malay

Fine-Tuning:
- English sentence → Model $\theta + \text{Head}$ → POS tags

Testing:
- Malay sentence → Model $\theta + \text{Head}$ → POS tags
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.

How Language-Neutral Are These Models?

- **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)**

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.

Aligning Representations in Multilingual Models

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

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.

..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.

One Solution: Adding Language-Specific Layers

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

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

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

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

(b) Mean Performance on XNLI and NER.

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

- Cross-Lingual Contrastive Learning (XLCO): Maximise sequence-level mutual information between parallel sentences
  
### Recall: Encoder, Decoder, Encoder-Decoder

<table>
<thead>
<tr>
<th><strong>Encoder-Only</strong></th>
<th><strong>Encoder-Decoder</strong></th>
<th><strong>Decoder-Only</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Typically trained on masked language modelling or similar</td>
<td>Trained on sequence-to-sequence data, or e.g. span corruption</td>
<td>Typically trained on autoregressive LM or similar</td>
</tr>
<tr>
<td>Outputs vectors/matrices</td>
<td>Outputs text</td>
<td>Outputs text</td>
</tr>
<tr>
<td>Fine-tuned for, e.g., classification tasks</td>
<td>Can be fine-tuned for various tasks</td>
<td>Often used with prompts and in-context learning</td>
</tr>
<tr>
<td>Includes BERT-type models</td>
<td>Includes (most) MT models</td>
<td>Includes GPT-type models</td>
</tr>
</tbody>
</table>
Multilingual Encoder-Decoder Models (Examples)

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

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

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

- **M2M-100**: Many-to-many parallel data training for 100 languages
  
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.
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
1 Motivation

2 Towards Multilingual MT

3 Multilingual Pre-Trained Models

4 Multilinguality in LLMs

5 Summary
### 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
Many “Large Language Models” (LLMs) are decoder-only
How to train a decoder-only model?

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

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

• 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 (!)
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
Multilingual MT & NLP

- English: 30.04%
- Chinese (Simplified): 16.2%
- French: 12.9%
- Portuguese: 4.9%
- Arabic: 4.6%
- Indic family: 4.4%
- Vietnamese: 2.7%
- Spanish: 10.8%
- Indonesian: 1.2%
- Catalan: 1.10%
- Basque: 0.15%
- Niger-Congo Family: 0.03%
- Code: 10.8%
State of Multilingual LLMs

% non-English pre-training

Nov '18         Nov '22

mBERT   XLM-R   mT5   mDeBERTaV3

BLOOM   XGLM

PaLM

LaMDA

Gopher

GLaM

OPT

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.

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


- 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

Summary

1 Motivation

2 Towards Multilingual MT

3 Multilingual Pre-Trained Models

4 Multilinguality in LLMs

5 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
Very broad view of multilingual pre-training, cross-lingual transfer, and related topics

Aim to give a high-level view/understanding

Many papers mentioned
→ Possible starting points for learning more

Thanks for listening even after end of term!