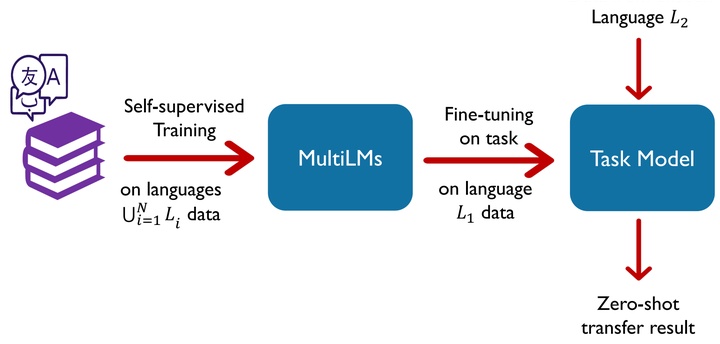
Discovering Language-neutral Sub-networks in Multilingual Language Models

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, Karl Aberer

EPFL

Multilingual Language Models (MultiLMs)

MultiLMs:
- Pre-trained jointly on raw data from multiple languages
- Fine-tuned for a task using a single high-resource language dataset

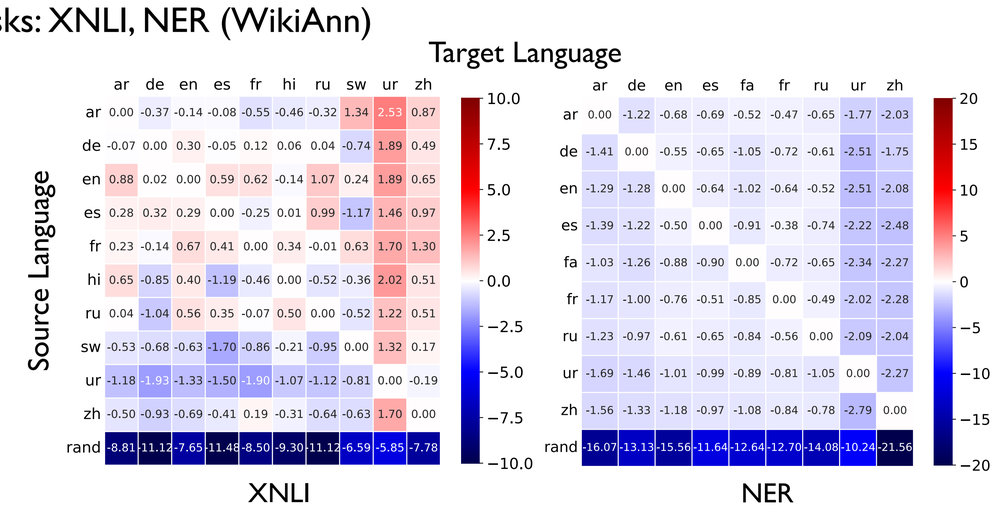


- They demonstrate promising cross-lingual transfer performance
- Language neutrality is considered a key facilitator of this performance
  - ➢ Shared representations that encode similar phenomena across languages

Do MultiLMs learn language-neutral parameters?

Sub-Networks Transfer Well Across Languages

Cross-lingual transfer evaluation on mBERT (50% sparsity):
- Languages: ar, de, en, fa, fr, hi, ru, ur, and zh
- Tasks: XNLI, NER (WikiAnn)

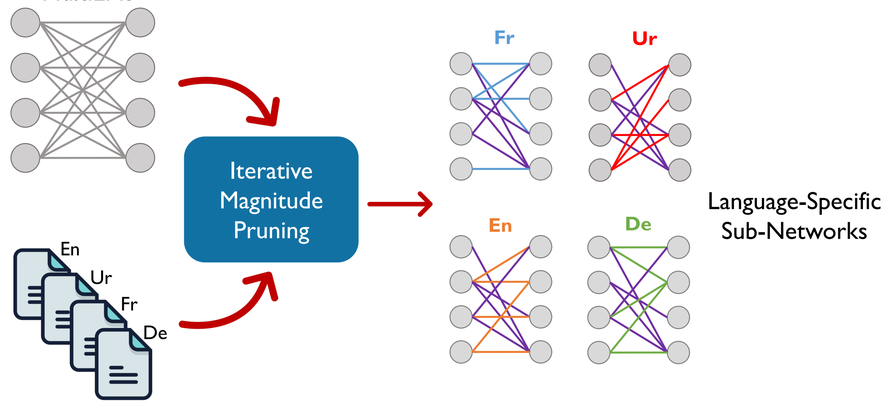


- High performance on cross-lingual transfer across languages (same task)

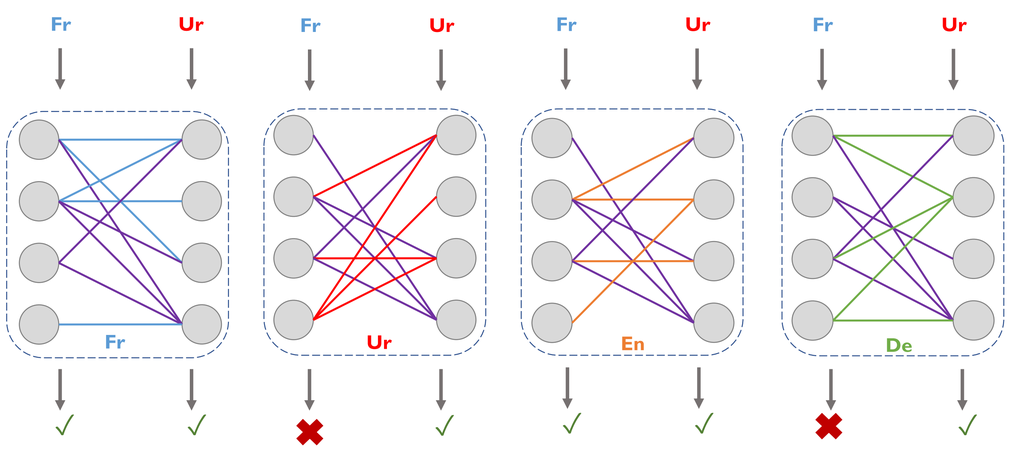
Language-specific sub-networks have shared multi-lingual components

Investigating Language Neutrality of MultiLMs

Hypothesis:
- Overlap between language-specific sub-networks indicates language neutrality



Extract sub-networks from MultiLMs:
- Using iterative magnitude pruning (Lottery Ticket Hypothesis [Frankle'19])
- Prune for individual language-task pairs



Evaluating transferability of extracted sub-networks on French (Fr) and Urdu (Ur).
Purple connections are shared in all sub-networks.

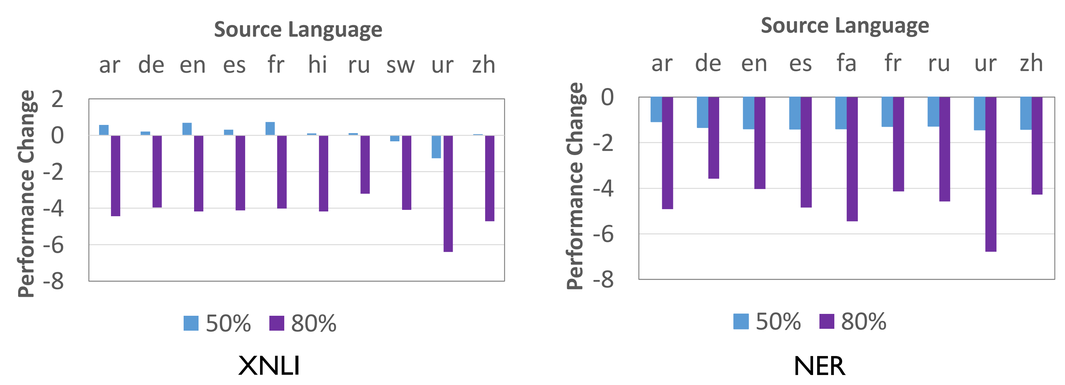
✓ → Transferred ✖ → Not Transferred

Evaluate the language neutrality degree of sub-networks:
- Transfer extracted sub-networks across languages
- Fine-tune them on other language-task pairs

Language neutrality → overlap between language-specific sub-networks

Sparsity Dampens Language Neutrality

Cross-lingual sub-network transfer degrades as the sub-networks get sparser
- Language-specific parameters are retained for the language for which a sub-network is discovered



Average of relative cross-lingual transfer performance drop for sub-networks with sparsity levels 50% and 80%. Relative performance change is computed as $\frac{1}{|L|-1}\sum_{t \in L \setminus s} \frac{acc(s,t)-acc(t,t)}{acc(t,t)}$ where s and t are source and target languages and L is the set of languages for each task.

Language-neutral parameters get pruned at higher sparsity

Sub-Networks Overlap Considerably

- NER sub-networks have the highest overlap
- MLM sub-networks have the lowest overlap
  - ➢ Upper layers are specialized to predict language-specific vocabularies
- Absolute overlap and cross-lingual performance are not correlated in a fine-grained manner

	MLM	XNLI	NER
Average Overlap	68%	85%	94%

The overlap among each task's sub-networks is high

Code

@negar.foroutan

negar-foroutan

Paper