Analyzing the Mono- and Cross-Lingual Pretraining Dynamics of Multilingual Language Models

Terra Blevins[1] Hila Gonen[1,2] Luke Zettlemoyer[1,2]
[1] University of Washington [2] Meta AI

What Do Multilingual Models Learn and When?

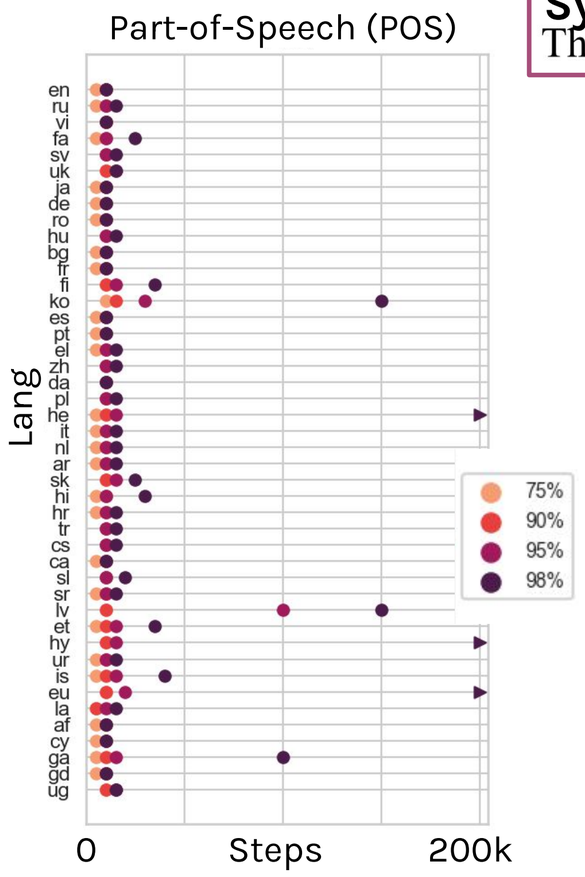
We replicate **XLM-R** and save **intermediate checkpoints** during pretraining to uncover **when** multilingual knowledge is learned.

Test $\text{XLM-R}_{\text{replica}}$ checkpoints on a range of linguistic tasks covering:

Syntax (amod) The quick brown fox jumps

Semantics The quick brown fox jumps / The fox is fast (Entails)

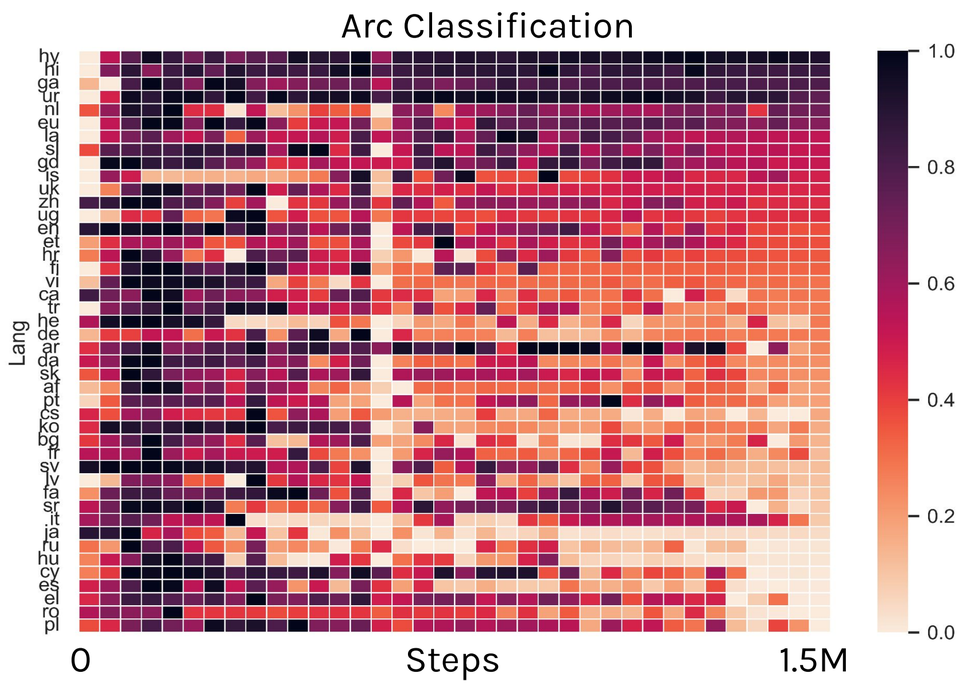
Word Alignment Le renard brun rapide saute / The quick brown fox jumps



XLM-R Learns Monolingual Information Early on

XLM-R learns 98% of POS knowledge for most languages by step 50k (3.3% of pretraining)

Learning and forgetting during pretraining for monolingual dependency arc classification



Key Findings

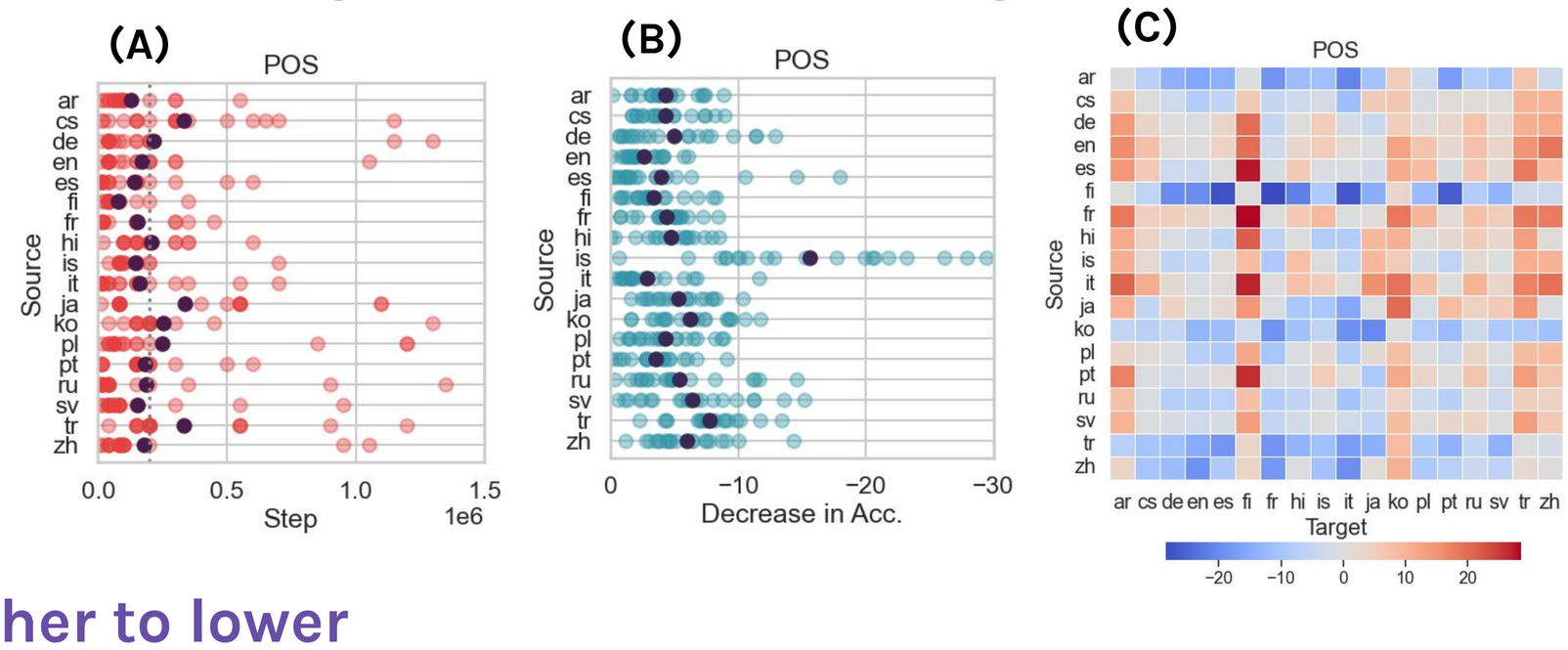
- XLM-R learns monolingual information first
- Cross-lingual knowledge is learned throughout pretraining
- Information moves to lower layers in the network

Cross-lingual Transfer Is Learned During the Entire Pretraining Process

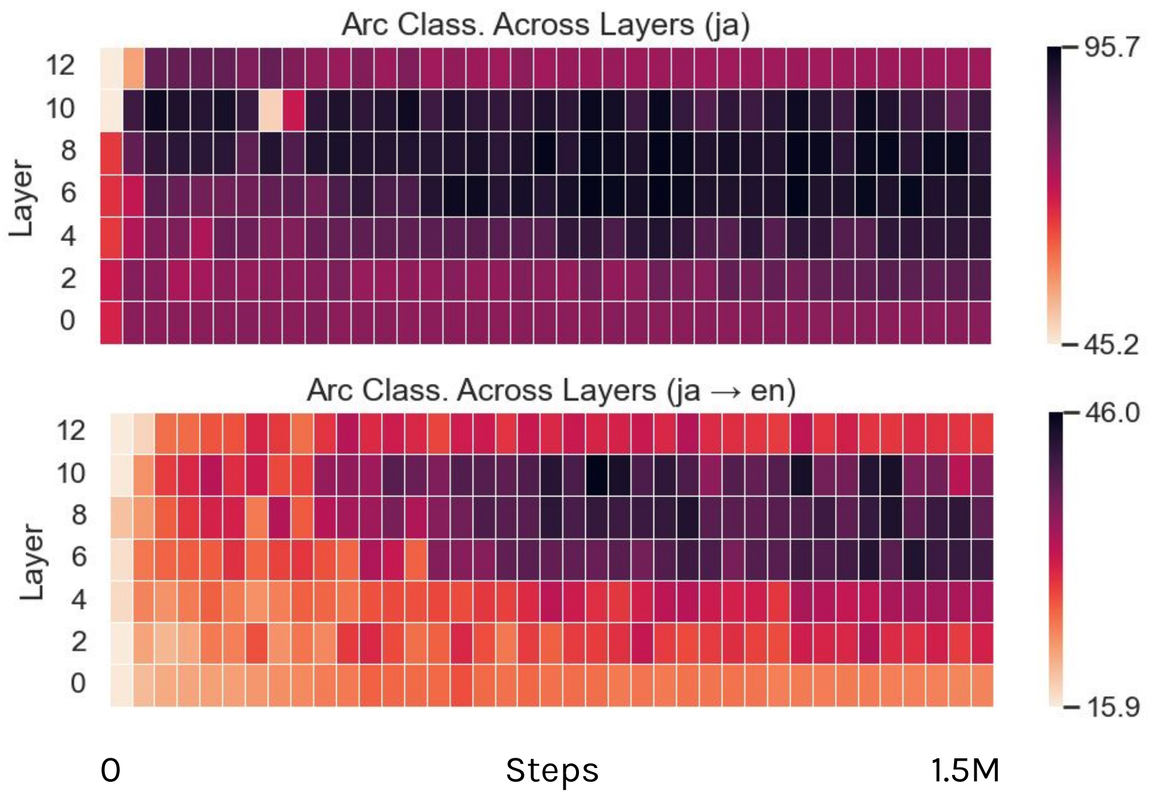
(A) Step in pretraining where XLM-R **learns** 98% of best cross-lingual score

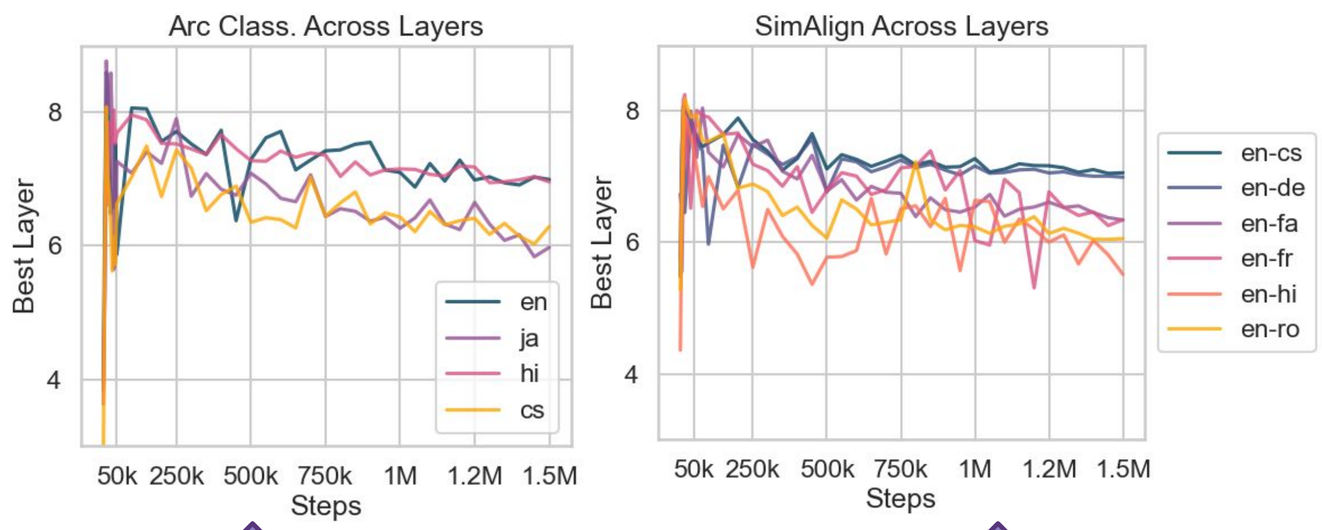
(B) XLM-R **forgets** cross-lingual information between the best and final checkpoints

(C) Cross-lingual transfer (and **when** languages transfer) is **asymmetric**



Information moves from higher to lower layers during pretraining





Monolingual and Cross-lingual

blvns.github.io

@terrablvns

More results + models in the paper!