Can Machine Translation Bridge Multilingual Pretraining and Cross-lingual Transfer Learning?

Shaoxiong Ji [1] Timothee Mickus [1] Vincent Segonne [2] Jörg Tiedemann [1]

[1] University of Helsinki [2] Universite Grenoble Alpes

firstname.lastname@helsinki.fi

Research questions

This paper investigates whether machine translation as a learning objective can improve performances on zero-shot cross-lingual transfer downstream tasks. We attempt to establish whether MT training objectives implicitly foster cross-lingual alignment:

(i) Do models (re)trained with the MT objective develop cross-lingual representations?

(ii) Do they generalize well on cross-lingual tasks?

(iii) Which factors impact their performances?

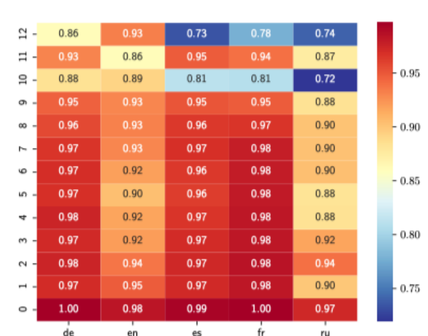
Findings

- MT (continued) training objectives do not favor the emergence of cross-lingual alignments more than LM objectives, based on the study on existing publicly available pretrained models.
- We provide evidence from similarity analyses and parameter-level investigations that this is due to separability, which is beneficial in MT but detrimental elsewhere.
- We conclude that MT encourages behavior that is not necessarily compatible with high performances in cross-lingual transfer learning.

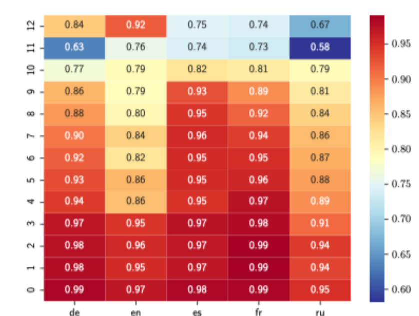
Quantitative performance

	Model	NC	XNLI	PAWS-X	QAM	QADSM	WPR	NER	POS
LM	mBERT	81.3	65.2	86.6	64.6	63.1	74.4	77.5	76.0
	XLM-R	**82.1**	**73.5**	88.9	67.4	**66.9**	**75.3**	**78.7**	**79.7**
	mBART	82.1	67.6	**89.2**	**67.8**	65.5	74.7	77.7	72.7
MT	NLLB 600M	76.0	68.3	73.4	61.5	63.9	73.7	54.2	71.4
CP	mBART m2o	80.4	65.9	85.6	63.9	63.9	73.7	61.5	70.8
	mBART o2m	65.4	48.1	81.7	58.4	62.7	73.2	55.1	55.7
	mBART m2m	78.3	60.2	87.2	63.2	62.8	73.7	71.9	69.7

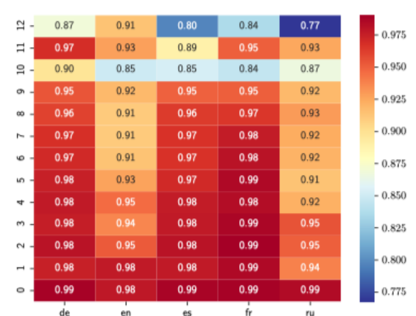
Table 1: Average performance on cross-lingual tasks. We use the base architecture for mBERT and XLM-R. mBART scores are derived from the 12-layer encoder.



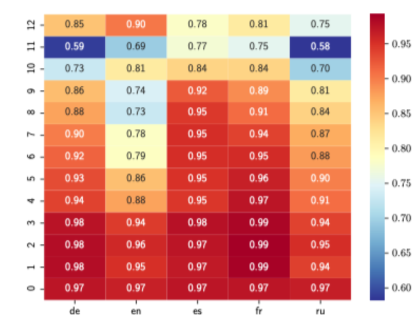
(a) mBART vs mBART m2m



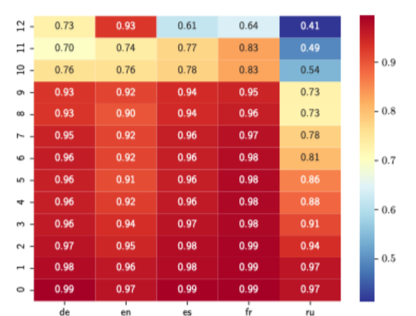
(b) XLM-R vs mBART m2m



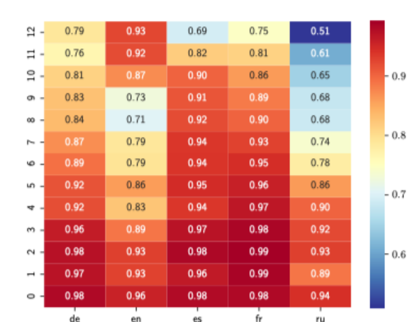
(c) mBART vs mBART m2o



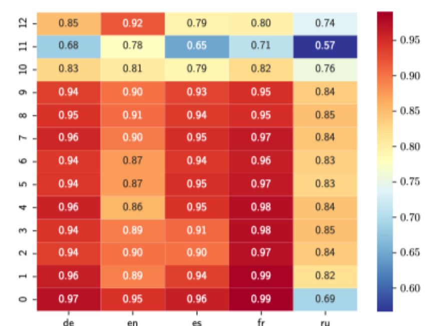
(d) XLM-R vs mBART m2o



(e) mBART vs mBART o2m

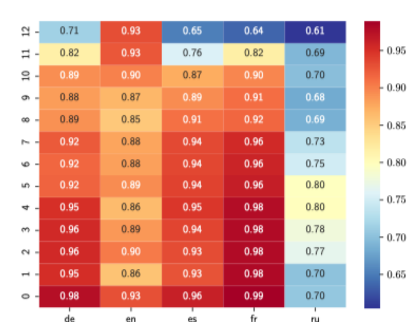


(f) XLM-R vs mBART o2m



(g) mBERT vs mBART m2m

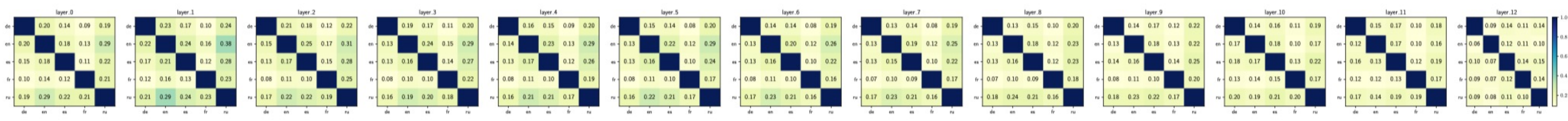
(h) mBERT vs mBART m2o



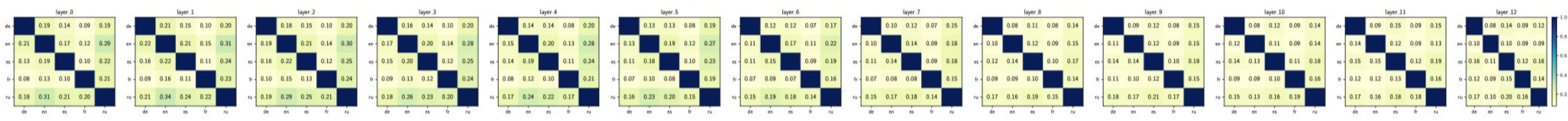
(i) mBERT vs mBART o2m

Figure 1: Representational similarity between mBART-based MT models and LMs

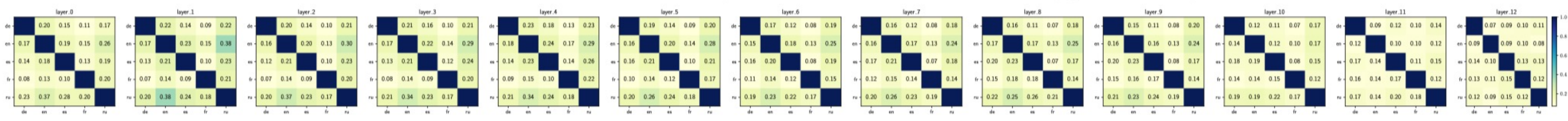
Representation similarity



(a) mBART (lower triangle) and mBART m2m (upper triangle)



(b) mBART m2o (lower triangle) and mBART o2m (upper triangle)



(c) mBERT (lower triangle) and XLM-R (upper triangle)

Figure 2: Representational similarity between different languages with representations learned by LMs and MT models

CP's effect on scaling

Model	K $\|\sigma\|$	K d	Q $\|\sigma\|$	Q d	V $\|\sigma\|$	V d	Out $\|\sigma\|$	Out d	FC up $\|\sigma\|$	FC up d	FC down $\|\sigma\|$	FC down d
mBART	44.76	–	44.85	–	53.73	–	53.45	–	90.25	–	99.63	–
mBART m2m	48.28	4.23	48.29	4.07	55.65	2.73	55.14	3.01	99.28	9.47	107.94	9.63
mBART m2o	48.34	4.23	48.35	4.06	56.19	2.95	55.73	2.99	101.06	11.19	109.71	11.18
mBART o2m	56.13	11.76	56.25	11.74	60.17	7.18	59.32	7.07	116.17	26.34	120.50	22.15

Table 2: SVD scaling effect for mBART and CP models; weight matrices from the 12th layer.

Cite as:
Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann.
Can machine translation bridge multilingual pretraining and cross-lingual transfer learning?
In Proceedings of LREC-COLING, 2024.

This work is part of the FoTran project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (agreement № 771113) and has received funding from the European Union's Horizon Europe research and innovation programme under Grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546]. The authors wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources, including Puhti, Mahti, and LUMI extreme scale access (MOOMIN and LumiNMT).