AlignFreeze: Navigating the Impact of Realignment on the Layers of Multilingual Models Across Diverse Languages

Steve Bakos, Félix Gaschi, David Guzmán, Riddhi More, Kelly Chutong Li, En-Shiun Annie Lee
felix@posos.fr

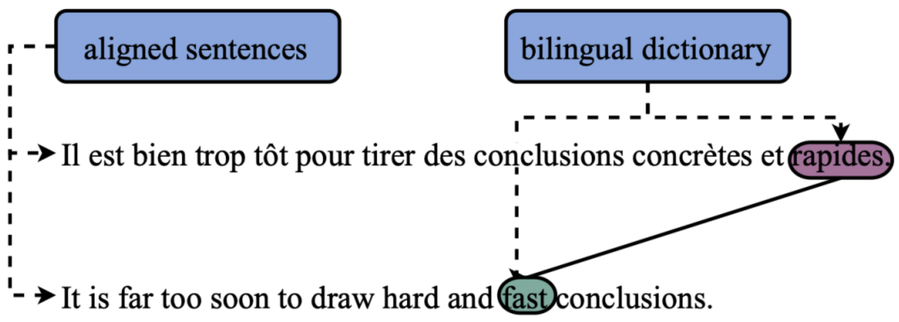
1. Motivation: mitigating realignment failures

- Multilingual Language Models (mLMs) like XLM-R and mBERT facilitate cross-lingual transfer
- Realignment techniques improve multilingual alignment but can degrade performance in some tasks and some languages

Realignment seemed to have a detrimental impact on some features learned during pre-training. But what features?

What are the layers for which realignment is the most detrimental and can we mitigate this effect?

2. Some context on realignment



From Gaschi et al. (2023)

The encoder-only multilingual model (mBERT, XLM-R, etc...) is trained with a contrastive loss to minimize the distance between translated words with respect to unrelated ones:

$$\mathcal{L}(\theta) = \frac{1}{2B} \sum_{h \in \mathcal{H}} \log \frac{\exp(\text{sim}(h, \text{aligned}(h))/T)}{\sum_{h' \in \mathcal{H}, h' \neq h} \exp(\text{sim}(h, h')/T)} \quad (1)$$

3. Contribution: A freezing approach

Two approaches and two baselines

- **Front-freezing**: freezes layers in the lower-half
- **Back-freezing**: freezes layers in the upper-half
- Baseline: **full realignment**, without freezing
- Baseline: **Simple fine-tuning**, no realignment

Depending on the results of each of the freezing method, we should understand better whether lower or upper layers are negatively impacted by realignment, i.e. whether low-level syntactic features or high-level semantic features are affected.

4. Full results

	PoS (34 lang.)			**NER** (34 lang.)			**NLI** (12 lang.)			**QA** (11 lang.)			Total (91)	
	acc.	#↓	#↑	acc.	#↓	#↑	acc.	#↓	#↑	F1	#↓	#↑	#↓	#↑
DistilMBERT														
Fine-tuning Only	73.8	-	-	82.5	-	-	60.1	-	-	38.1	-	-	-	-
Full realignment	77.6	0	31	84.7	3	21	61.6	3	5	39.3	2	5	8	62
AlignFreeze (front)	76.2	0	34	84.0	1	21	61.6	1	8	37.4	4	2	6	65
AlignFreeze (back)	77.4	0	30	83.7	4	17	61.9	1	6	39.1	2	5	7	58
mBERT														
Fine-tuning Only	77.0	-	-	85.7	-	-	66.3	-	-	57.1	-	-	-	-
Full realignment	79.6	1	32	86.4	19	4	67.4	0	8	52.9	11	0	31	44
AlignFreeze (front)	79.2	0	32	**86.7**	1	6	67.7	0	10	55.3	9	0	10	48
AlignFreeze (back)	79.3	1	30	86.5	12	6	67.5	0	10	53.7	11	0	24	46
XLM-R Base														
Fine-tuning Only	80.9	-	-	84.9	-	-	**73.9**	-	-	**61.2**	-	-	-	-
Full realignment	81.3	1	11	85.3	8	8	73.2	8	0	59.4	10	0	27	19
AlignFreeze (front)	**81.7**	0	18	84.8	11	4	73.6	6	0	59.1	10	0	27	22
AlignFreeze (back)	80.9	7	4	84.9	13	7	72.9	11	0	58.0	11	0	42	11
Total of #↓ and #↑ by task		/102			/102			/36			/33		/273	
Full realignment	-	2	74	-	30	33	-	11	13	-	6	6	64	125
AlignFreeze (front)	-	0	84	-	13	31	-	7	18	-	9	2	43	135
AlignFreeze (back)	-	8	64	-	29	30	-	12	16	-	11	10	73	115

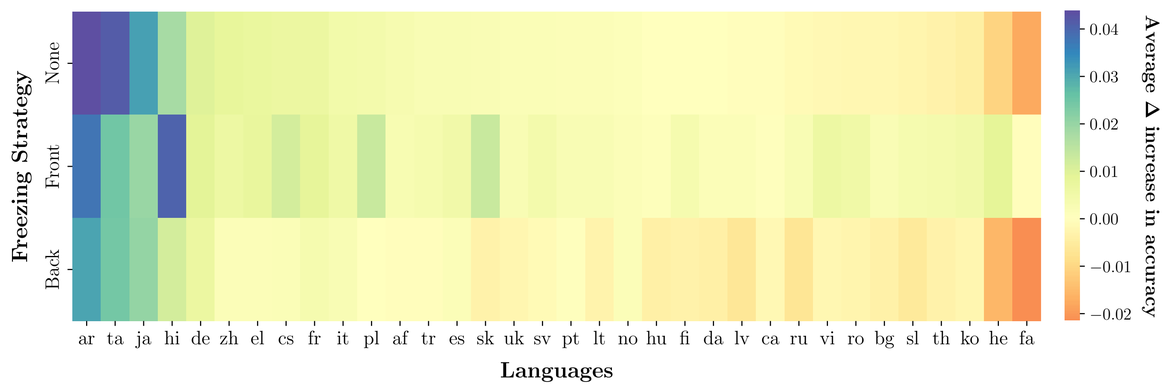
#↑: number of target languages for which the realignment accuracy is one standard deviation above the simple fine-tuning baseline

#↓: number of target languages for which the realignment accuracy is one standard deviation below the simple fine-tuning baseline

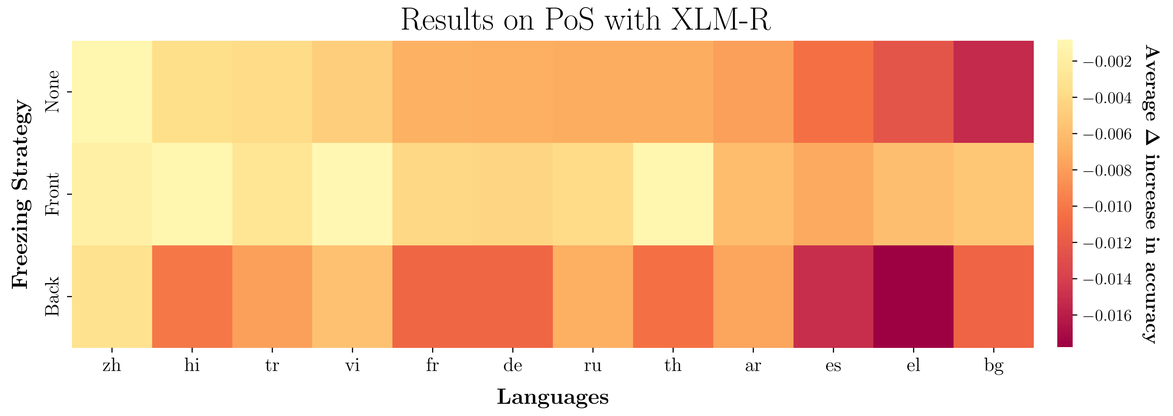
Findings

- **Full realignment fails in many cases** (as already shown in previous literature)
- **AlignFreeze (front) mitigates the failures of full realignment to some extent**

5. Results across languages



Results on PoS with XLM-R



Results on NLI with XLM-R

Realignment impacts the whole model for all languages, but it is the most detrimental to the lower layers.

6. Conclusion

- AlignFreeze shows that realignment has a particularly detrimental impact on lower layers
- New lead for improving realignment: preserve low-level features
- No one-size-fits-all solution; results vary across tasks, languages, and models.
- Further research needed to optimize freezing strategies and analyze language-specific effects.