

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

Authors: Steve Bakos, Félix Gaschi, David Guzmán, Riddhi More, Kelly Chutong Li, En-Shiun Annie Lee

Institutions: Ontario Tech University, University of Toronto, SAS Posos, France

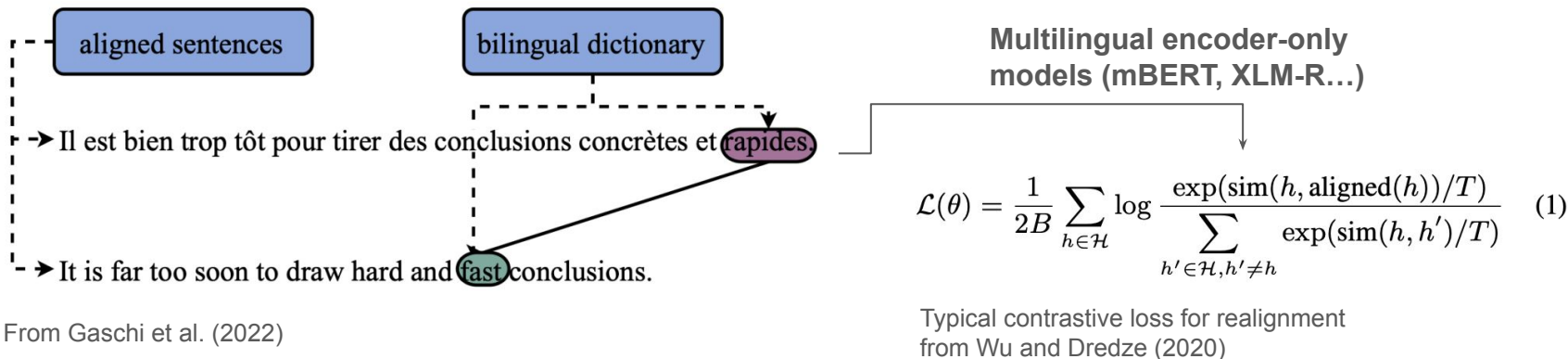
Introduction: investigating realignment failure

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

Realignment has adverse effect on some features learned during pre-training.

AlignFreeze: A method that freezes half of the model during realignment to better understand on which layers realignment is more impactful

Methodology: freeze layer during realignment



From Gaschi et al. (2022)

- **Front-Freezing:** Freezes lower-half layers while realigning the upper layers.
- **Back-Freezing:** Freezes upper-half layers while realigning the lower layers.
- **Full realignment (baseline):** All layers are realigned.

Results: AlignFreeze mitigates realignment failures

	# failures	# successes
Full realignment	64	125
Front freezing	43	135
Back freezing	73	115

Aggregated results over all tasks, models, languages

Failures: realignment provide degradation over simple fine-tuning >1std

Successes: realignment provide improvement over simple fine-tuning >1std

Observations:

- Realignment often fails (agree with Wu & Dredze 2020)
- Realignment works better for smaller models and low-level tasks (agree with Gaschi et al. 2023)
- Front freezing particularly improves performances for PoS tagging with distant languages

Front freezing mitigates realignment failures.

→ **Realignment is detrimental to lower layers, i.e. to lower-level encoded features**

Conclusions

- **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.