Towards Best Practices for Training Multilingual Dense Retrieval Models.

Xinyu Zhang (Crystina), Kelechi Ogueji, Xueguang Ma, Jimmy Lin

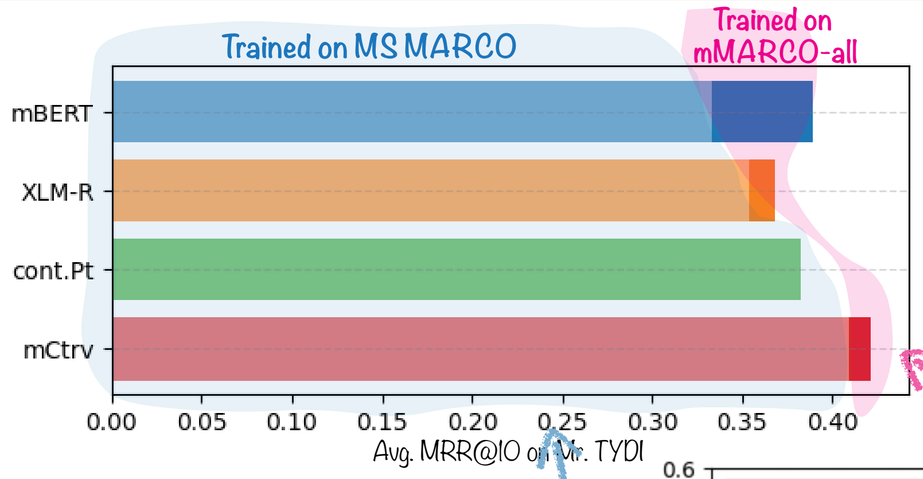
ACM Transactions on Information Systems

SIGIR 2024 Washington, D.C.

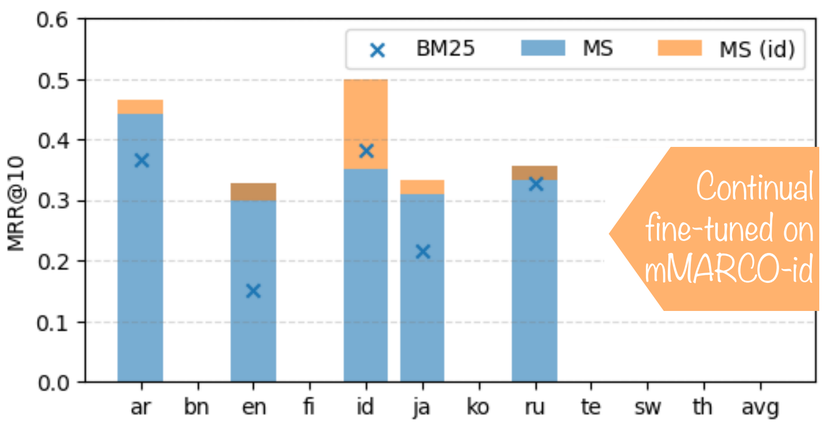
BEST PRACTICE ALONG {PT - Cont. PT - Pre-FT - FT} PIPELINE

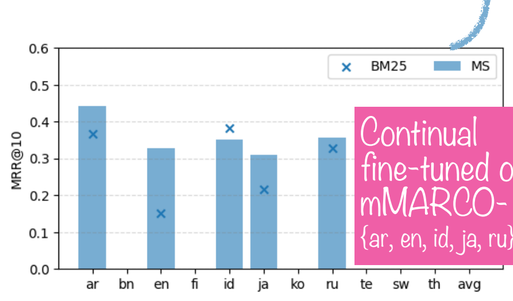
PT

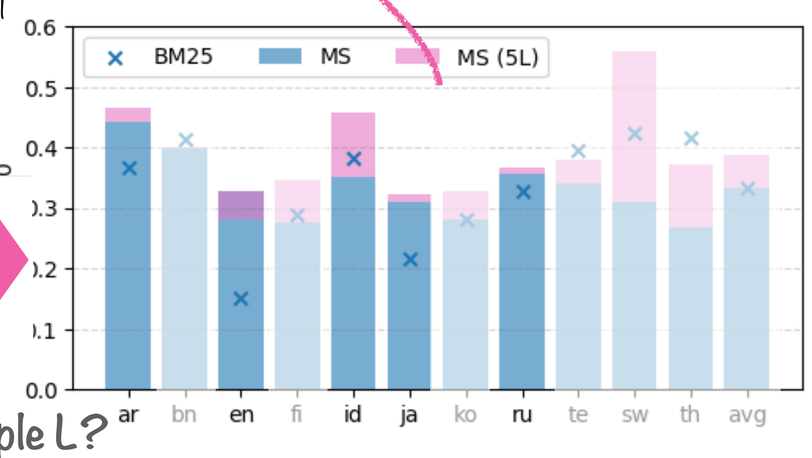
Representatives from:
- different pLM class;
- Enhanced Language knowledge
- Enhanced Task knowledge



Cont. PT



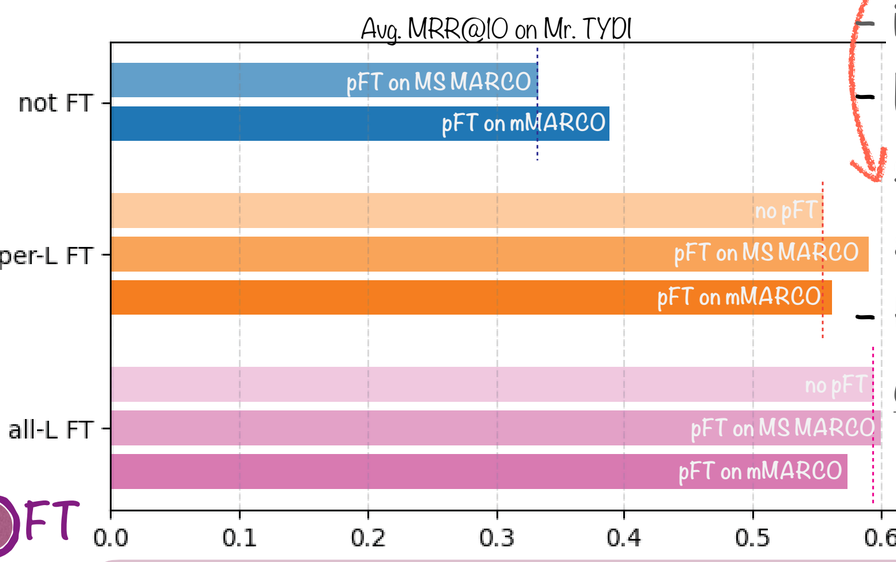




- Pre-FT on single or multiple L?

  in-L single > out-L multiple > out-L single

FT

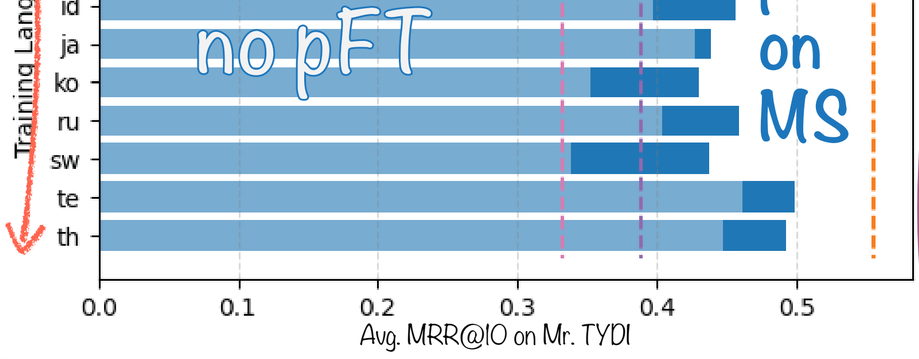


- in-domain in-L data is good ofc!
- Does the multilingual pFT data still helpful when in-domain data available? Not Really!
- How about multilingual in-domain data? Still Helpful!

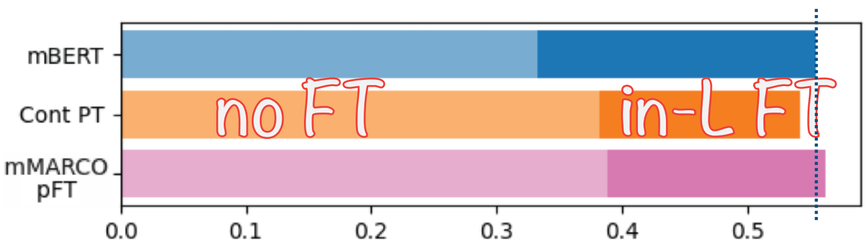
Have data in target L

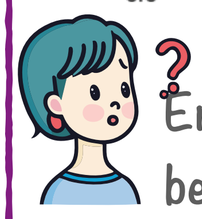
- How about in-domain but out-L data?

  in-L > out-L > pFT

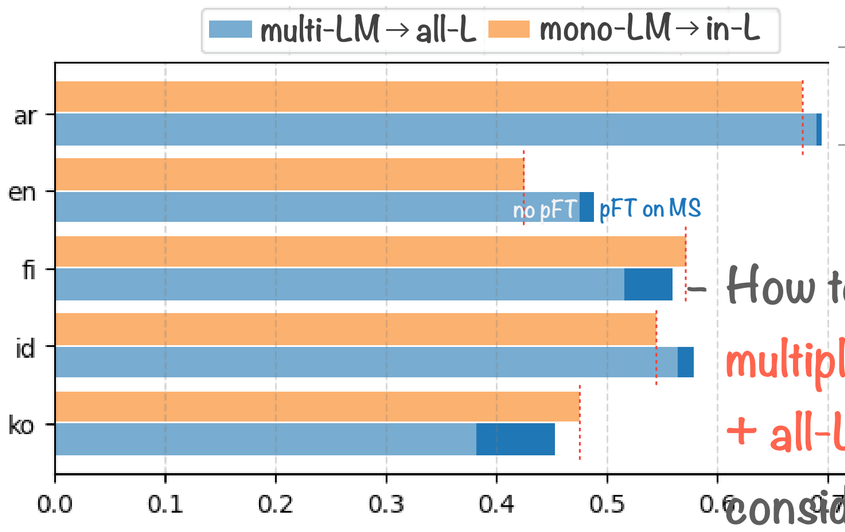


No data in target L; Have data in non-target L





Enhanced Language knowledge before FT all becomes not helpful when in-domain in-L data is available.
(e.g. Cont. PT, mMARCO)



LM type	Langauge Speciality	Utilize non-target-L Data
mono-L	✓	✗
multi-L	✗	✓

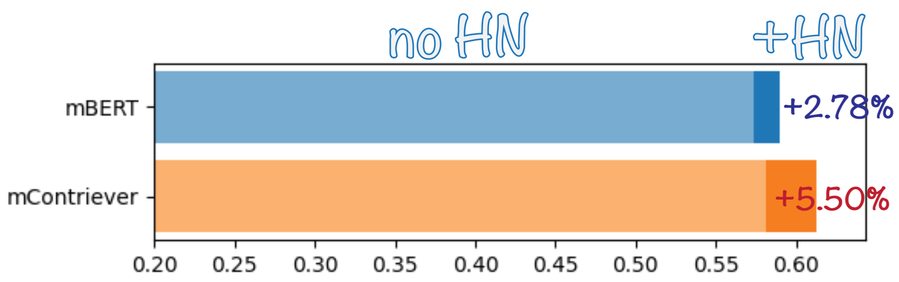
- How to choose? With FT-data from multiple languages available, {multi-LM + all-L} wins in general (especially when considering the cost)

Have data in target and non-target L: {mono-LM + in-L} OR {multi-LM + all-L}?

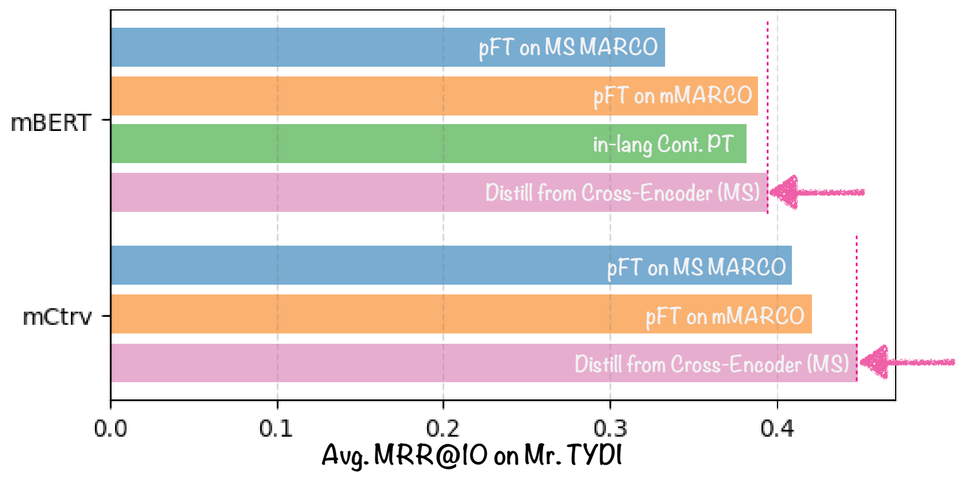
Has Data In Target Domain

AUXILIARIES EXPERIMENTS

- Add Hard Negatives? Yes!



- Distillation? Yes!



Paper Link