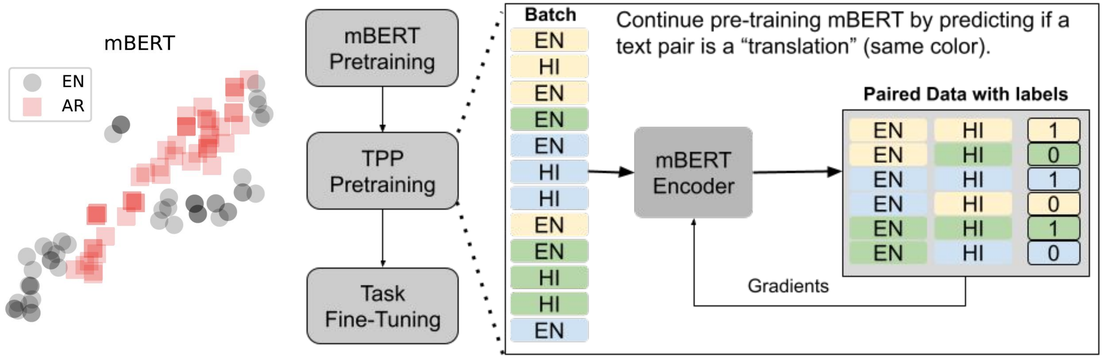
Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction

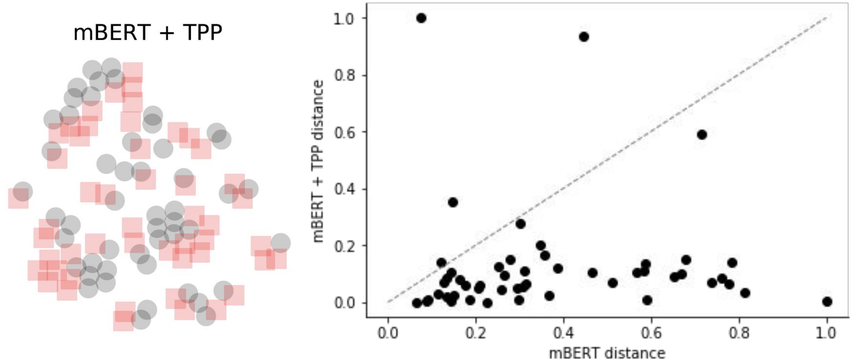
Shubhanshu Mishra, Aria Haghighi | Twitter, Inc.

2021 The 7th Workshop on Noisy User-generated Text (W-NUT)

Code: github.com/twitter-research/multilingual-alignment-tpp

Is mBERT aligned? No. Can we align it to improve zero-shot transfer on social media text? Yes.





- Misalignment of Language Models → lower zero-shot transfer capabilities.
- Significant accuracy drop for orthographically different languages.
- Availability of translation pairs of varying quality can align Language Models.

Translation Datasets (Size)

Lang pair	Tatoeba	Wikimatrix	Wikidata
en-ar	28K	773K	1.6M
en-ja	220K	480K	509K
en-hi	11K	134K	77K

Translations from wikidata descriptions and labels (**WD**) **[NEW]**

Translations mined from Wikipedia using Cross Lingual Model (**WM**)

Human written translations (**TT**)

Translation Pair Prediction (TPP) Setup

- **mBERT:** Baseline
- **+TPP (ONE)**: Single pair training.
- **+TPP (BP)**: Consecutive pair training on best two dataset.
- **+TPP (ALL)**: All language pair training.

Downstream Zero Shot Evaluation Setup

- Fine-tune on only English dataset for the task
- **Hypothesis:** Alignment helps zero-shot transfer.
- This assumption may fail when translation of task does not exist:
  - E.g. abuse in one language not translatable in other language.
- NER and Sentiment dataset are based on Tweets, UD POS is included to check performance in standard domain.

Downstream performance

	Hindi		Japanese		Arabic	
NER	F_1	Δ%	F_1	Δ%	F_1	Δ%
mBERT	21.1	0.0	16.5	0.0	32.1	0.0
+TPP (ONE)	**24.3**	15.2	**29.9**	81.4	**39.4**	22.8
+TPP (ALL)	23.2	10.3	27.4	66.4	38.5	19.9
Sentiment	F_1	Δ%	F_1	Δ%	F_1	Δ%
mBERT	31.7	0.0	55.0	0.0	51.5	0.0
+TPP (ONE)	**32.7**	3.0	66.4	20.6	58.3	13.2
+TPP (ALL)	32.4	2.3	**67.7**	23.1	**58.5**	13.7
UD POS	acc.	Δ%	acc.	Δ%	acc.	Δ%
mBERT	67.4	0.0	52.7	0.0	64.0	0.0
+TPP (ONE)	**71.5**	6.0	**57.6**	9.2	**67.1**	4.8
+TPP (ALL)	66.4	-1.5	52.7	0.1	65.0	1.5

- **NER**: 37% relative improvement in F1.
- **Sentiment**: 12% relative improvement in F1.
- **UD POS**: 6.7% relative improvement in accuracy.

Impact of Translation Quality (see paper for details)

- **Tatoeba** is likely to be the most accurate as it is manually curated.
- **Wikidata** is likely to be higher quality for HI (low resource)
- **Wikimatrix** is auto generated hence likely to perform worse on low-resource languages compared to AR and JA (high resource).

Conclusion

- TPP is a simple way to align any encoder.
- Don't expect embeddings or models trained on all languages data to share information across orthographically different languages
- Task type impacts transfer:
  - **Good**: Syntactic tasks (NER, POS)
  - **OK**: Semantic tasks (Sentiment, Abuse).
- Our results are promising given the lack of social media bitext corpus.
- Our downstream setup can serve as a benchmark to evaluate multilingual performance on social media text.