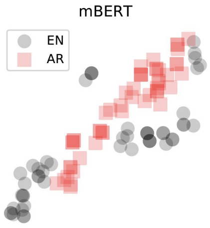
Improved Multilingual Language Model Pretraining for Social Media Text via Translation Pair Prediction

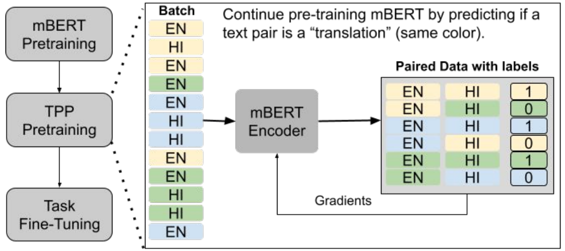
Shubhanshu Mishra, Aria Haghighi | Twitter, Inc.

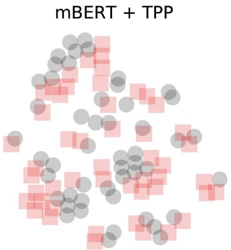
2021 The 7th Workshop on Noisy User-generated Text (W-NUT)

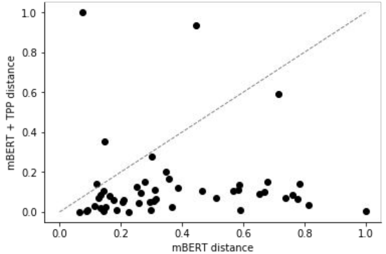
Code: github.com/twitter-research/multilingual-alignment-tpp

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 diff. 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

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).

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

Conclusion

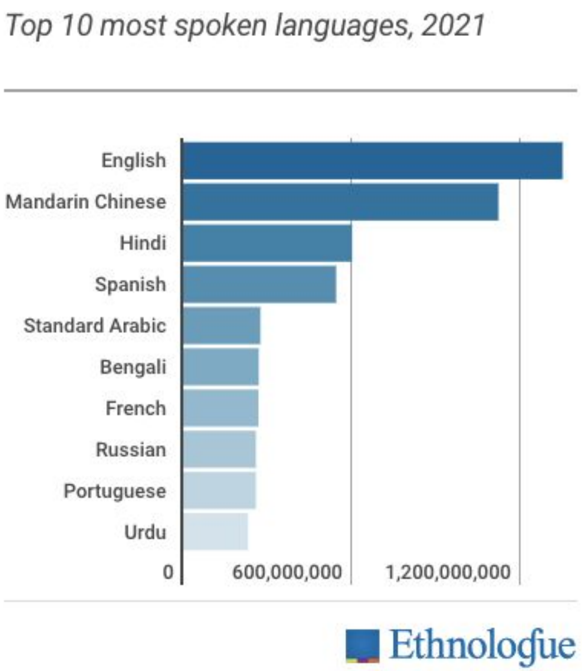
- TPP is simple way to align any encoder.
- Don't expect embeddings or models trained on all languages data to share information across orthographically different languages
- Task type impacts transfer:
  - **Good**: Syntactic tasks (NER, POS)
  - **OK**: Semantic tasks (Sentiment, Abuse).
- Our results 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.

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

Shubhanshu Mishra, and Aria Haghighi
Twitter, Inc.

2021 The 7th Workshop on Noisy User-generated Text (W-NUT)
Code: github.com/twitter-research/multilingual-alignment-tpp

Why multilingual models?



Source: https://www.ethnologue.com/guides/ethnologue200

Languages		Regions	Participation			Active editors					Edits		Usage	Content
Code ⇒ Project Main Page	Language ⇒ Wikipedia article		Speakers in millions (log scale) (?) Editors per million speakers (5+ edits)	Prim.+Sec. Speakers M=millions k=thousands	Editors (5+) per million speakers	Months since 3 or more active editors	5+ edits p/month (3m avg)	100+ edits p/month (3m avg)	Admins	Bots	Bot edits	Human edits by unreg. users	Views per hour	Article count
Σ	All languages	AF AS EU NA SA OC CL W												
en	English	AF AS EU NA OC		1121 M	27		30684	3445	1274	312	9%	31%	4,858,539	5,779,516
ceb	Cebuano	AS		20 M	1		26	2	4	60	99%	19%	1,311	5,379,752
sv	Swedish	EU		10 M	64		641	101	66	40	57%	20%	53,206	3,761,531
de	German	EU		132 M	41		5395	900	198	374	10%	20%	726,852	2,254,737
fr	French	AF AS EU NA OC SA		285 M	17		4864	790	161	107	19%	21%	461,591	2,069,464
nl	Dutch	EU SA		28 M	42		1185	214	45	269	38%	19%	97,322	1,953,504
ru	Russian	AS EU		264 M	12		3188	518	87	84	17%	25%	634,782	1,518,909
es	Spanish	AF AS EU NA SA		513 M	8		4135	544	71	36	17%	37%	417,439	1,496,759
it	Italian	EU		68 M	35		2355	398	109	173	29%	32%	270,709	1,489,914
pl	Polish	EU		43 M	29		1256	237	106	68	34%	19%	185,774	1,313,943

Source: https://stats.wikimedia.org/EN/Sitemap.htm#comparisons

Source:
https://tatoeba.org/eng/sentences/show/657403

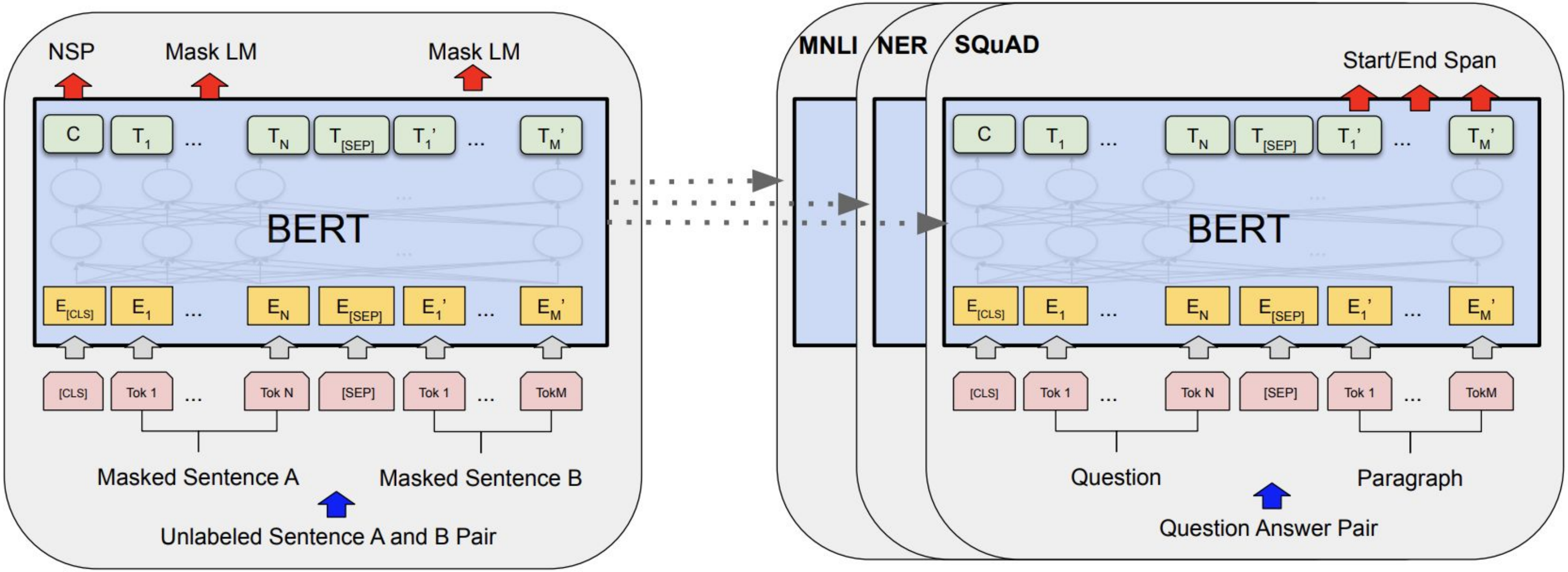
Motivation: Multilingual NER

NER trained on tweets using Multilingual Word Embeddings and BiLSTM

Language Testing Dataset	English CoNLL-03	German CoNLL-03	Dutch CoNLL-02	Spanish CoNLL-02	French xLIME	Italian xLIME	Turkish JRC	Hindi SEAS	Arabic CS-18
Lookup	36.6	22.8	36.8	29.7	15.6	23.3	22.9	**20.4**	16.7
Mono Training	40.2	35.5	39.4	27.4	27.7	**29.3**	24.8	11.8	**22.8**
Mul Training	38.3	36.6	43.2	29.1	26.4	28.9	28.0	9.8	14.0
Mono Training + WikiANN	**47.2**	**41.2**	**55.4**	37.6	30.3	28.4	27.8	14.0	21.9
Mul Training + WikiANN	43.2	39.6	52.8	**44.0**	**32.6**	25.4	**28.6**	8.3	11.3

Table 1: Entity-Level Micro-Average F1-scores for the PERSON, LOCATION and ORGANIZATION types

Table Source: Ramy Eskander, Peter Martigny, Shubhanshu Mishra. Multilingual Named Entity Recognition in Tweets using Wikidata in WeCNLP 2020



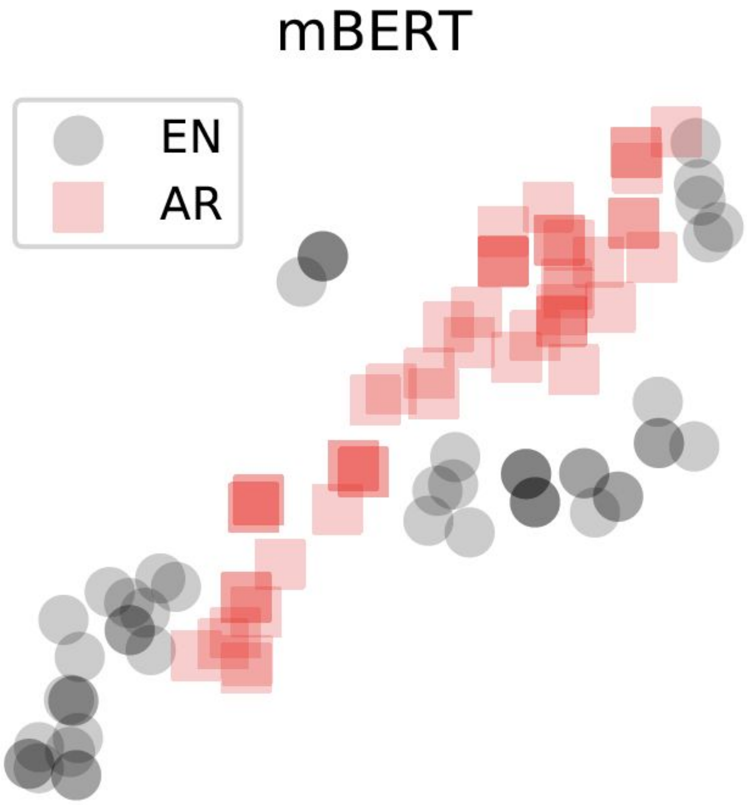
Pre-training

Fine-Tuning

Source: [1810.04805] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

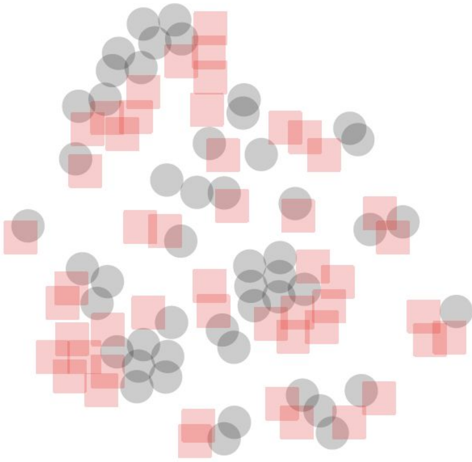
We use the Multilingual BERT as an encoder for representing the text and tokens.

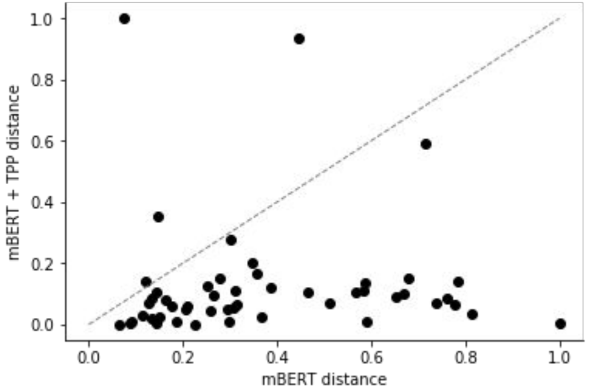
Is mBERT aligned?
No.



Can we align it to improve zero-shot transfer on social media text?
Yes.

mBERT + TPP



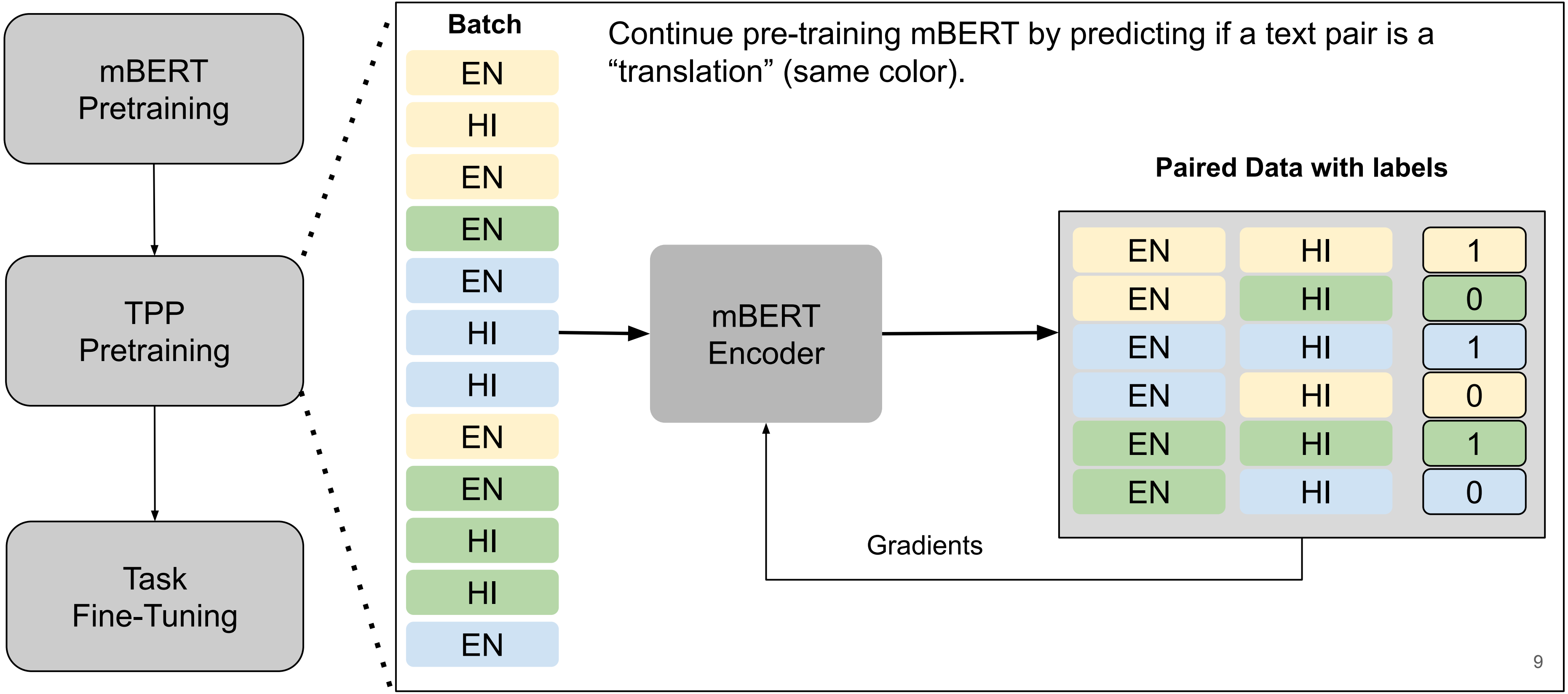


Misalignment of Language Models → lower zero-shot transfer capabilities.

Significant accuracy drop for orthographically diff. languages.

Availability of translation pairs of varying quality can align Language Models.

Translation Pair Prediction → New Pretraining task



Continue pre-training mBERT by predicting if a text pair is a "translation" (same color).

Translation Datasets (Size)

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

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

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

Human written translations (**TT**)

Wikidata Translation Pairs

natural language processing (Q30642)

field of computer science and linguistics
NLP

▾ In more languages
Configure

Language	Label	Description
English	natural language processing	field of computer science and linguistics
Spanish	procesamiento de lenguajes naturales	subdisciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional
Traditional Chinese	自然語言處理	No description defined
Chinese	自然语言处理	以通过语音输入文字为例，自然语言处理是用计算机来处理、理解以及运用人类语言。

For each wikidata items with label and description in languages part of translation pair, e.g. English (en) and Hindi (hi), create sentences as follows:

```
Source[en]  : Label[en] + “ ” + Description[en]
Target[hi]  : Label[hi] + “ ” + Description[hi]
```

Translation Pair Prediction (TPP) Setup

- **mBERT:** Baseline
- **+TPP (ONE):** Single pair training.
  - The language pair data comes from either Tatoeba (TT), Wikimatrix (WM), or Wikidata (WD).
- **+TPP (BP):** Consecutive pair training on best two dataset.
  - mBERT → TPP(TT) → TT(WM).
- **+TPP (ALL):** All language pair training.
  - mBERT → TPP(TT-AR + TT-HI + TT-JA)
  - This model can give us a good trade-off for model serving and improved accuracy.

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 Zero Shot Evaluation Setup

	Hindi		Japanese		Arabic	
NER	F_1	$\Delta\%$	F_1	$\Delta\%$	F_1	$\Delta\%$
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	$\Delta\%$	F_1	$\Delta\%$	F_1	$\Delta\%$
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.	$\Delta\%$	acc.	$\Delta\%$	acc.	$\Delta\%$
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.

Performance using various translation pairs (NER)

	Hindi		Japanese		Arabic	
NER	F_1	$\Delta\%$	F_1	$\Delta\%$	F_1	$\Delta\%$
mBERT	21.1	0.0	16.5	0.0	32.1	0.0
+TPP (TT)	23.1	9.6	27.8	68.6	36.3	13.2
+TPP (WD)	22.4	6.3	26.5	60.8	36.9	15.0
+TPP (WM)	21.6	2.6	27.7	68.3	38.3	19.3
+TPP (BP)	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

Performance using various translation pairs (Sentiment)

	Hindi		Japanese		Arabic	
Sentiment	F_1	$\Delta\%$	F_1	$\Delta\%$	F_1	$\Delta\%$
mBERT	31.7	0.0	55.0	0.0	51.5	0.0
+TPP (TT)	31.8	0.3	62.4	13.5	58.3	13.2
+TPP (WD)	30.8	-2.9	50.2	-8.7	53.0	3.0
+TPP (WM)	32.7	3.0	63.2	14.8	54.7	6.4
+TPP (BP)	32.0	0.9	66.4	20.6	55.3	7.5
+TPP (ALL)	32.4	2.3	67.7	23.1	58.5	13.7

Performance using various translation pairs (UD POS)

	Hindi		Japanese		Arabic	
UD POS	acc.	Δ%	acc.	Δ%	acc.	Δ%
mBERT	67.4	0.0	52.7	0.0	64.0	0.0
+TPP (TT)	65.1	-3.5	54.0	2.4	66.7	4.1
+TPP (WD)	70.5	4.5	53.0	0.5	66.4	3.7
+TPP (WM)	70.4	4.3	54.4	3.1	65.4	2.2
+TPP (BP)	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

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 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.

Thank You!

Questions @TheShubhanshu and @aria42

Code and experiment details at:
https://github.com/twitter-research/multilingual-alignment-tpp