Analysis → MuRIL: Multilingual Representations for Indian Languages (link)

1. There are multiple Indian languages in the state of the Art Multilingual transformers , but the representation of Indian languages in their vocabulary and training data remains very low .
2. MuRIL is a multilingual LM designed particularly for IN languages trained on vast quantities of IN text corpora. It is deliberately supplemented monolingual text corpora with document pairings that are both translated and transliterated, serving as supervised cross-lingual signals in training.
3. The first is the traditional Masked Language Modeling (MLM) goal (predict the masked token given a sentence) , which relies solely on monolingual text data (unsupervised). The second goal is Translation Language Modeling (TLM)(predict token on parallel language space) , which makes use of parallel data (supervised) .
4. Assamese (as), Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Kashmiri (ks), Malayalam (ml), Marathi (mr), Nepali (ne), Oriya (or), Punjabi (pa), Sanskrit (sa), Sindhi (sd), Tamil (ta), Telugu (te), and Urdu (ur) are among the IN languages plus English are the languages on which it is trained .
5. There were three streams of data which they used to fetch the data .
    a. Monolingual Data in the respective language.
    b. Translated Data , where each pair has a native tongue phrase and its English translation. Then , they employ an in-house translation system to convert the monolingual corpora (both Common Crawl and Wikipedia) to English. The model is trained using both the original and translated documents as parallel instances.
    c. Transliterated Data -They had a dataset comprising 10,000 phrase pairings for 12 IN languages (bn, gu, hi, kn, ml, mr, pa, ta, te, ur). Each pair consists of a phrase written in native script and its carefully romanized transliteration.They transliterate Wikipedia corpora of all IN languages to Latin using the indic-trans library .
6. The distribution of tokens per language is extremely unbalanced. As a result, data smoothing is required to ensure that all languages' representations are accurately reflected . So they are upscaled by using the formula
    a. mi = (max j∈L nj /ni)^(1−α)(α=0.33)
7. It is observed that mBERT has a greater fertility ratio than MuRIL. Following are the possible reasons for it :
    a. mBERT vocabulary has relatively limited representation of IN languages
    b. The vocabulary does not take transliterated terms into consideration.
8. A greater fertility ratio translates to a larger number of sub-words per word, resulting in a loss of semantic meaning retention.
9. Pre-Training Details - We utilise the MLM and TLM goals to pre-train a BERT base encoder model. We train for 1M steps (with 50k warm-up steps and a linear decay) with a maximum sequence length of 512 and a global batch size of 4096.

Model	PANX	UDPOS	XNLI	Tatoeba	XQuAD	MLQA	TyDiQA-GoldP	Avg.
	F1	**F1**	**Acc.**	**Acc.**	**F1/EM**	**F1/EM**	**F1/EM**	
mBERT	58.0	71.2	66.8	18.4	71.2/58.2	65.3/51.2	63.1/51.7	59.1
MuRIL	**77.6**	**75.0**	**74.1**	**25.2**	**79.1/65.6**	**73.8/58.8**	**75.4/59.3**	**68.6**

Table 1: *Results for MuRIL and mBERT on XTREME (IN).* We observe that MuRIL significantly outperforms mBERT on all the datasets in XTREME. Note that here we present the average performance on test sets for IN languages only that MuRIL currently supports. Please refer to Section 3 for more details.

Model	PANX	UDPOS	XNLI	Tatoeba	Avg.
	F1	**F1**	**Acc.**	**Acc.**	
mBERT	14.2	28.2	39.2	2.7	21.1
MuRIL	**57.7**	**62.1**	**64.7**	**11.0**	**48.9**

Table 2: *Results for MuRIL and mBERT on XTREME (IN-tr).* We transliterate IN language test sets (native → Latin) and present the average performance across all transliterated test sets. Please refer to Section 3 for more details.

Reason to read → The reason why I read this is to make the approaches language agnostic at least for the set of Indian Languages .
Directions to use → link