Exploiting Language Relatedness for Low Web-Resource Language Model Adaptation: An Indic Languages Study

Yash Khemchandani[1], Sarvesh Mehtani[1], Vaidehi Patil[1], Abhijeet Awasthi[1], Partha Talukdar[2], Sunita Sarawagi[1]
[1]Indian Institute of Technology Bombay, India [2]Google Research, India
{yashkhem, smehtani}@cse.iitb.ac.in, vaidehipatil@ee.iitb.ac.in

Challenges

- Recent research in multilingual language models (LM) holds promise for low web-resource languages (LRL) as multilingual models can enable transfer of supervision from high resource languages to LRLs
- Incorporating a new language in an LM still remains a challenge, particularly for languages with limited corpora and in unseen scripts
- The current paradigm for training Multilingual LM requires text corpora in the languages of interest, usually in large volumes. However, such text corpora is often available in limited quantities for LRLs
- Relatedness among languages in a language family may be exploited to overcome some of the corpora limitations of LRLs

Motivation

Transliteration

When transliterated to the same script, overlapping words across related languages serve as anchors in multilingual pre-training

Hindi	करण दिल्ली जा रहा है
Punjabi	ਕਰਨ ਦਿੱਲੀ ਜਾ ਰਿਹਾ ਹੈ
Punjabi in Devanagari script	करन दिली जा रिहा है

LRL	Hindi	English
Punjabi	**25.5**	7.5
Gujarati	**23.3**	4.5
Bengali	**10.9**	5.5

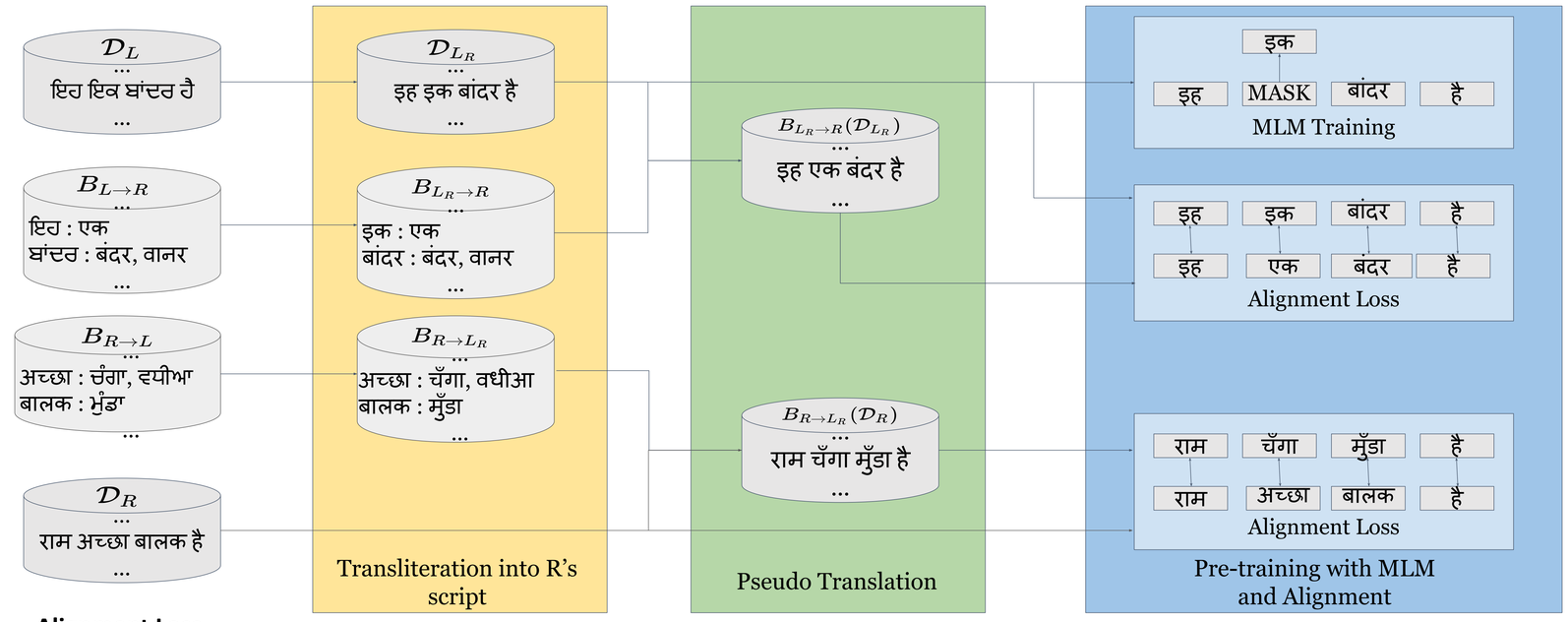
Pseudo-Translation

Exploits various syntactic similarities such as Sentence-Object-Verb order

Hindi Sentence	Pseudo-Translation to Punjabi	Gold Translation
यह एक प्यारी बिल्ली है	ਇਹ ਇਕ ਪਿਆਰਾ ਬਿੱਲੀ ਹੈ	ਇਹ ਇੱਕ ਪਿਆਰੀ ਬਿੱਲੀ ਹੈ

Hindi	यह	एक	प्यारी	बिल्ली
Punjabi	ਇਹ	ਇਕ	ਪਿਆਰਾ	ਬਿੱਲੀ

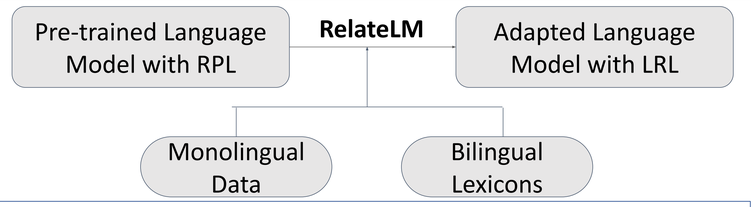
RelateLM architecture



Alignment Loss

MSE loss that attempts to bring the embeddings of corresponding tokens closer

Contributions

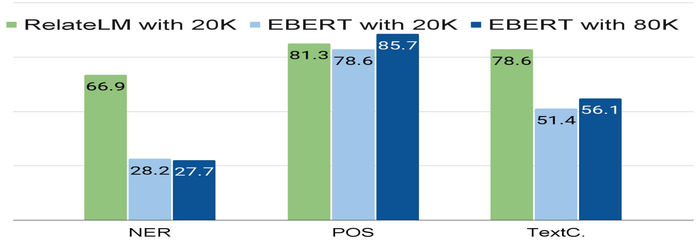


- We propose **RelateLM** : a language model which exploits relatedness between a Low Web-Resource language (LRL) and a Related Prominent Language (RPL)
- Given a LM pre-trained on an RPL, we propose an effective method to incorporate a "related" LRL

Results

LRL Adaptation	Prominent Language	Punjabi		
		NER	POS	TextC.
EBERT	en	19.4	48.6	33.6
RelateLM - PseudoT	en	38.6	58.1	54.7
EBERT	hi	28.2	78.6	51.4
RelateLM - PseudoT	hi	65.1	77.3	76.1
RelateLM	hi	**66.9**	**81.3**	**78.6**

LRL adaptation	Prominent Language	NER	POS	TextC.
Oriya				
RelateLM-PseudoT	en	14.2	72.1	63.2
RelateLM	en	16.4	74.1	62.7
EBERT	hi	10.8	71.7	53.1
RelateLM-PseudoT	hi	22.7	74.7	76.5
RelateLM	hi	**24.7**	**75.2**	**76.7**



Conclusion

- EBERT performs much better when RPL is Hindi than when RPL is English due to language similarity within a language family.
- RelateLM - PseudoT performs much better when RPL is Hindi than when RPL is English due to increased token overlap when RPL is a Closely Related Language.
- RelateLM - PseudoT performs better than EBERT both when RPL is Hindi and English due to target LRL being in the same script as RPL.
- RelateLM almost always performs the best showing the power of exploiting language relatedness through Transliteration (increased token overlap) and Word-Level Alignment of LRL with RPL