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

Authors: Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, Sunita Sarawagi

Mrunmayee Amshekar, Sunil Dhaka

Indian Institute of Technology, Kanpur

mrunmayee22@iitk.ac.in, sunild@iitk.ac.in

Computational Linguistics for Indian Languages

November 11, 2022

Table of Contents

Background

- There exist language models pre-trained for multilingual settings (e.g. mBERT - 104 languages)
- Mutlilingual LMs fine-tuned for various tasks, they transfer knowledge from resource rich to low resource languages
- Mutlilingual LMs require large monolingual corpora
- Recent development for LRL - train a light-weight adaptive layer(keeping the full model fixed), exploit overlapping tokens to learn embeddings of the LRL
- general purpose methods are there that do not exploit the specific relatedness of languages within the same family

Background

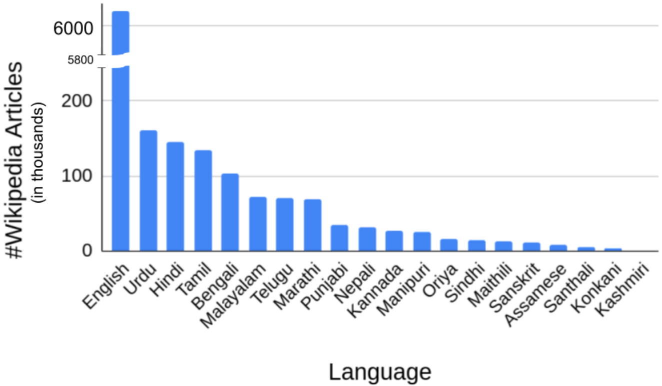

Figure: Number of Wikipedia articles in English and top Indian languages

RelateLM I

- Relatedness - translation and parallel data from related IA languages - previously used to improve NMT (Goyal et al., 2020)
- RelateLM exploits the relatedness between LRLs and a related prominent language (RPL) in **language models**. RPL for Indic languages - **Hindi**
- Relatedness against two dimensions is considered
  - script
  - sentence structure
- Why Hindi?
  - For 3 Indic languages - overlapping tokens with Hindi range between $11 - 26$ % as against $< 8\%$ in English, that is mostly numbers and names
  - Syntax-level similarities between languages allows us to enrich data using bilingual dictionaries

RelateLM II

- Subject-Object-Verb (SOV) order of Indo-Aryan family
- Brahmic family simplifies rule-based transliteration libraries for any language combination

- Authors demonstrate how Oriya and Assamese can be adapted for mBERT using RelateML
- Also compare how using RelateML affects various tasks in different languages through different LMs

Approach

- RelateLM supplement existing model M with LRL through an existing related prominent language in M
- 3-step approach
    - Transliteration to RPL script - using IndicTrans library (Bhat et al, 2015)
    - Pseudo-translation - parallel tagged corpus may not be available for LRLs. 2 factors facilitate pseudo-translation
        - Word level bilingual dictionaries(RPL-LRL) or crowd sourcing (cheaper than dataset creation by experts)
        - Common syntactic properties e.g. word order
    - Adaptation - following transliteration and translation, a union of 2 pseudo parallel corpora (RPL to LRL and LRL to RPL) using alignment loss
        - align corresponding word embeddings to bring multilingual embeddings in different languages closer

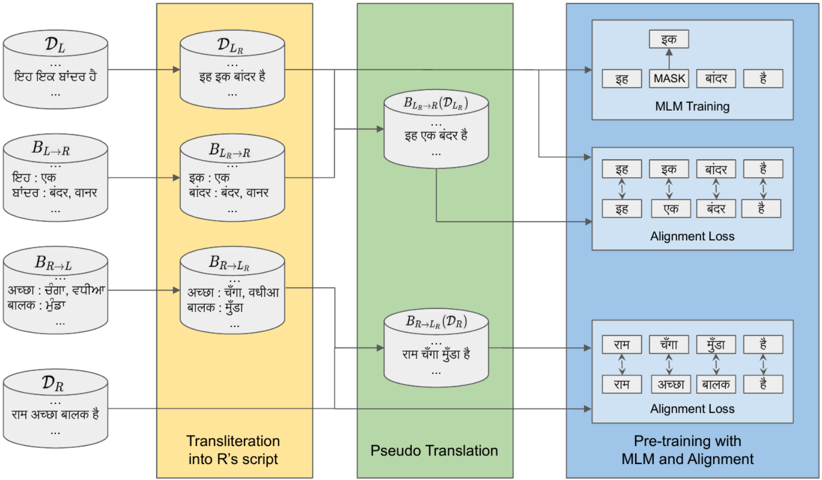

Figure: Process of RelateML

Experiment

- Evaluate whether RelateLM can extend Oriya (unseen script) and Assamese (seen script) in mBERT
- Adapt LM trained on RPL to LRL using RelateLM
- Data sourced from English and Hindi Wikipedia articles. A comparison of English and Hindi as the RPLs
- 3 methods compared for five languages for the tasks - NER, POS and Text classification
  - EBERT
  - RelateLM without pseudo translation
  - mBERT - if it is trained for the language
- after that have to decide whether to include or exclude such observations

Findings

- Transliterating LRLs to Hindi provides gains over transliterating to English and also EBERT
- Gains were more significant for Oriya than Assamese
- Pseudo-translation to Hindi - added gains

Findings

LRL Adaptation	Prominent Language	Punjabi NER	Punjabi POS	Punjabi TextC.	Gujarati NER	Gujarati POS	Gujarati TextC.	Bengali NER	Bengali POS	Bengali TextC.
mBERT	-	41.7	86.3	64.2	39.8	87.8	65.8	70.8	83.4	75.9
EBERT (Wang et al., 2020)	en	19.4	48.6	33.6	14.5	56.6	37.8	31.2	50.7	32.7
RelateLM−PseudoT	en	38.6	58.1	54.7	15.3	58.5	57.2	**68.8**	59.8	58.6
EBERT (Wang et al., 2020)	hi	28.2	78.6	51.4	14.8	69.0	48.1	34.0	**73.2**	45.6
RelateLM−PseudoT	hi	65.1	77.3	76.1	39.6	80.2	79.1	56.3	69.9	77.5
RelateLM	hi	**66.9**	**81.3**	**78.6**	**39.7**	**82.3**	**79.8**	57.3	71.7	**78.7**

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 (Wang et al., 2020)	hi	10.8	71.7	53.1
RelateLM−PseudoT	hi	22.7	74.7	76.5
RelateLM	hi	**24.7**	**75.2**	**76.7**
Assamese				
RelateLM−PseudoT	en	-	78.2	74.8
RelateLM	en	-	77.4	74.7
EBERT (Wang et al., 2020)	hi	-	71.9	78.6
RelateLM−PseudoT	hi	-	**79.4**	79.8
RelateLM	hi	-	79.3	**80.2**

Findings

- Compared to bilingual language models HI-BERT and BERT
- tasks - NER, POS, Dictionary lookup (first, max entry, weighted, root weighted)
- higher gains when LRL transliterated to Hindi than English
- lower gains for Bengali - distinct phonology, TB influence

References

https://arxiv.org/abs/2106.03958

The End

Questions? Comments?