Multilingual Models in the Wild: Performance of Multilingual Models on Code-Mixed Tasks

22 November 2019

Ahmet Üstün, Prajit Dhar, Gosse Bouma

a.ustun@rug.nl, p.dhar@rug.nl, g.bouma@rug.nl

What is code-mixed/code-switching

- Code-mixing is usage of different languages within a single conversation context in an alternative manner.
- It is a prominent phenomenon in social media.

What is code-mixed/code-switching

1. <Original>Yesssssss, eindelijk de #seizoensfinale van #Familie! **Let's kill June**
   <English>Yesssssss, finally the season final of Family! Let's kill June
2. <Original>**Basic** kan dus heel **stylish** zijn.
   <English>So basic can be very stylish.3
3. <Original>Er wordt weer lustig er op los **geframed** door de NOSover #brexit.
   <English>Again at the NOS they are freely framing about #brexit

Multilingual Models

- Massively multilingual sentence representation models, trained on with/out parallel corpora
- Multilingual Bert & XLM (Cross-Lingual Language Model Pretraining)
  - Trained single BERT model on +100 languages from Wikipedia with a shared wordpiece/byte-pair vocabulary
    - https://github.com/google-research/bert/blob/master/multilingual.md
    - https://github.com/facebookresearch/XLM
- LASER (Language-Agnostic Sentence Representation)
  - A encoder-decoder based, supervised multilingual sentence representation model with shared byte-pair vocabularies.
    - https://github.com/facebookresearch/LASER

Corresponding Literature

- How multilingual is Multilingual BERT
  https://arxiv.org/pdf/1906.01502.pdf
  - An experiments on Hindi-English POS tagging task
  - Further analysis **is needed !!!**

	Corrected	Transliterated
Train on monolingual HI+EN		
M-BERT	86.59	50.41
Ball and Garrette (2018)	—	77.40
Train on code-switched HI/EN		
M-BERT	90.56	85.64
Bhat et al. (2018)	—	90.53

Table 6: M-BERT's POS accuracy on the code-switched Hindi/English dataset from Bhat et al. (2018), on script-corrected and original (transliterated) tokens, and comparisons to existing work on code-switch POS.

Possible Research Direction

- Tasks:
  - Language identification (Code-Mixing detection) on Code-Mixed data
  - Part-of-Speech (POS) Tagging
  - Sentiment analysis
  - Universal dependency parsing
  - ...
- Research directions:
  - Comparison of (multilingual) models on a task
  - Comparison of performance across different tasks
  - Comparison of performance across different language pairs
  - Examining different adaptation techniques for code-mixed tasks
  - ...

university of groningen

Example Data

- Hindi - English POS data
  https://github.com/UniversalDependencies/UD_Hindi_English-HIENCS/tree/master
- Hindi-English and Spanish-English sentiment analysis data
  https://ritual-uh.github.io/sentimix2020/
- Turkish-English code-mixing data
  https://www.aclweb.org/anthology/W18-6115.pdf
- ...