Analysis of Negative Interference in Multilingual Models with Code-Switched Data

DIWAN ANUJ JITENDRA	170070005
ASHISH MITTAL	204057001
YASH SHARMA	17D070059
ADITYA VAVRE	170050089

On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment

(https://www.aclweb.org/anthology/2020.emnlp-main.359.pdf)

- Training a neural network with multiple languages is shown to hurt performance on some languages

- **Analyzing negative interference**
  For a language pair, pre-train a single bilingual model and two monolingual models.

- The models are then compared in **within-language monolingual setting** and **zero-shot cross-lingual transfer setting.**

Our Proposal: Analyze Negative Interference in Code-Switching

- Our goal is to analyze the effect of negative interference (during pretraining of the LM on different sets of languages) on downstream code-switching tasks.
- The downstream code-switching tasks we focus on:
    a. Sentiment Analysis
    b. Natural Language Inference

- Our contributions:
    a. Run a number of experiments to explore the above phenomenon
    b. Make a code-switching raw text corpus for pretraining. (We finally didn't use this for pretraining due to its small size, but we created a reasonably good corpus)

Methodology

- All experiments are run using the XLM language model using https://github.com/facebookresearch/XLM and https://github.com/iedwardwangi/MetaAdapter as our basis.
- All experiments have the following high-level procedure:
  a. BPE-tokenize raw text corpus for LM pretraining
  b. Pretrain XLM model using the above raw text corpus
  c. Finetune the pretrained model on the downstream code-switching tasks (downstream data is also tokenized using the same BPE vocab)
  d. Report val and test acc and F1 score

Dataset Description

- **Pretraining:** Various sources including
  a. **English:** 54 million sentences Wikipedia dump, WMT Common Crawl, WMT NewsCommonCrawl.
  b. **Hindi:** 67 million sentences from Ai4Bharat, IITBombay Hindi Monolingual Corpus (Kunchukuttan et al.,2018)

- **Finetuning:** The GLUECoS code-switching dataset is used. We use the following 2 downstream tasks:
  a. Sentiment Analysis
     17500 training instances
  b. Natural Language Inference
     1500 training instances

Creating a HiEn code-switched text corpus

- There is no large raw text dataset for Hindi-English text.
- Therefore, we sourced the text from various public datasets (Hi-En tweet dataset, hinglishNorm, LINCE, PHINC)
- Tweets were processed using the Python package tweet-preprocessor, our own CoNLL parser, and cleaned using clean-text.
- Most of the data contained Hindi written in the Latin script. Thus, we use a Hindi wordlist available in the GLUECoS repository and Microsoft Azure’s Transliterate API to transliterate Latin-script Hindi words to Devanagari.

Description of each Model – Pre-Training method

1. **XLM**: Original architecture and MLM objective function of XLM (with a smaller number of layers (6) and a smaller hidden dimension (512))
2. **XLM_LN:** Original Architecture + language-specific linear layers
3. **XLM_FFN:** Original Architecture + language-specific feedforward layers
4. **XLM_ATTN:** Original Architecture + language-specific attention layers
5. **XLM_ADPT:** Original Architecture + residual adapter layers (Rebuffi et al. 2017, Houlsby el al. 2019)
6. **XLM-R:** Large model (12 layers, 1024 hidden dimension) pretrained on 17 languages.

Pretraining Experimental Details

- We have trained Bilingual Models for 300 epochs, with each epoch having 200000 iterations with batch-size 32 on a single GPU. The training took around a week.
- Monolingual models were trained for 200 epochs with each epoch having 200000 iterations with batch-size 32 on a single GPU. The training took around 3 days..
- The perplexities for each of the model was less than 10, and we used 80000 BPE tokens.
- We earlier tried to use a smaller dataset (~1 million sentences) but couldn't get the perplexity down to less than 400.
- Checkpoints are available at https://drive.google.com/drive/folders/1iODmKqIG9i9RKM8mqy0kvVzEH90to0Sg?usp=sharing

Experimental Results

Model	Pre	F1	Acc
XLM	EN, HI	0.65	65.39
XLM_LN	EN, HI	0.65	65.83
XLM_FFN	EN, HI	0.64	64.32
XLM_ATTN	EN, HI	0.64	64.64
XLM_ADPT	EN, HI	0.65	64.88
XLM	**EN**	**0.69**	**69.16**
XLM	**HI**	**0.67**	**67.97**
XLM-R	17 langs	0.2	45.63

Sentiment Analysis

Model	Pre	F1	Acc
XLM	EN, HI	0.41	42.46
XLM_LN	EN, HI	0.34	43.83
XLM_FFN	EN, HI	0.46	46.57
XLM_ATTN	EN, HI	0.59	53.42
XLM_ADPT	**EN, HI**	**0.69**	**52.73**
XLM	EN	0.44	46.57
XLM	HI	0.46	44.52
XLM-R	**17 langs**	**0.69**	**52.73**

Natural Language Inference

Discussion

1. For both the downstream tasks, monolingual pretraining (with just En or just Hi) performed better than the vanilla bilingual models (En+Hi).
2. For Sentiment analysis, monolingual models outperforms even enhancements in the XLM model.
3. For NLI, enhancements outperform monolingual models.

Bilingual pretrained models interfere with code-switched language tasks

Related Work

- Negative Interference in general multitask models: *Overcoming Negative Transfer: A Survey* (Zhang et. al. 2020) presents a survey of negative transfer for many multitask learning problems. *Gradient Surgery for Multi-Task Learning* (Yu et. al. 2020) describes how to fix conflicting gradients in multitask learning.
- Aligning multilingual embeddings: Papers like MUSE i.e. *Word Translation Without Parallel Data* (Conneau et. al. 2017), *Multilingual Alignment of Contextual Word Representations* (Cao et. al. 2020), , etc. acknowledge that multilingual models often have unaligned cross-lingual representations and special methods need to be developed to align them.
- Code switching: Survey papers like *A Survey of Code-switched Speech and Language Processing* explore the phenomenon of code-switching.

Conclusion & Future Work

- We explore the hitherto unexplored phenomenon of negative interference during pre-training, on code-switched data.
- We develop a code-switched raw text corpus and run experiments on two downstream code-switching tasks.
- We find that interference does occur even for code-switching data and negatively affects downstream performance.

Future work

- Exploring use of data augmentation for code-switching (like GCM).
- Exploring other approaches like gradient surgery to solve negative interference and conflicting gradients in multilingual models

References

1. *On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment* ; Zirui Wang, Zachary C. Lipton, Yulia Tsvetkov (2020)
2. *Cross-lingual Language Model Pretraining* ; Guillaume Lample, Alexis Conneau (2019)
3. *GLUECoS : An Evaluation Benchmark for Code-Switched NLP* ; Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, Monojit Choudhury (2020)
4. *Rebuffi, Sylvestre-Alvise, Hakan Bilen, and Andrea Vedaldi. "Learning multiple visual domains with residual adapters." arXiv preprint arXiv:1705.08045 (2017).*
5. *Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." International Conference on Machine Learning. PMLR, 2019.*