A Semi-supervised Approach to Generate the Code-Mixed Text using Pre-trained Encoder and Transfer Learning

This folder contains the code-mixed datasets created using the approach discussed in Section 3 of our paper:

License:

Citing the paper:

Please cite our paper if you use any of the datasets in your research work.

```
@inproceedings{gupta-etal-2020-semi,
      title = "A Semi-supervised Approach to Generate the Code-Mixed Text using
Pre-trained Encoder and Transfer Learning",
    author = "Gupta, Deepak  and
      Ekbal, Asif  and
      Bhattacharyya, Pushpak",
    booktitle = "Findings of the Association for Computational Linguistics: EMNLP 2020",
    month = nov,
    year = "2020",
    address = "Online",
    publisher = "Association for Computational Linguistics",
    url = "https://www.aclweb.org/anthology/2020.findings-emnlp.206",
    doi = "10.18653/v1/2020.findings-emnlp.206",
    pages = "2267--2280",
     abstract = "Code-mixing, the interleaving of two or more languages within a sentence
or discourse is ubiquitous in multilingual societies. The lack of code-mixed training data
is one of the major concerns for the development of end-to-end neural network-based
models to be deployed for a variety of natural language processing (NLP) applications.
A potential solution is to either manually create or crowd-source the code-mixed labelled
data for the task at hand, but that requires much human efforts and often not feasible
because of the language specific diversity in the code-mixed text. To circumvent the
data scarcity issue, we propose an effective deep learning approach for automatically
generating the code-mixed text from English to multiple languages without any parallel
data. In order to train the neural network, we create synthetic code-mixed texts from the available parallel corpus by modelling various linguistic properties of code-mixing. Our codemixed text generator is built upon the encoder-decoder framework, where the encoder is augmented with the linguistic and task-agnostic features obtained from the transformer based language model. We also transfer the knowledge from a neural machine translation (NMT) to warm-start the training of code-mixed generator. Experimental results and in-depth analysis show the effectiveness of our proposed code-mixed text generation on eight diverse language pairs.",
}

@inproceedings{gupta-etal-2018-uncovering,
    title = "Uncovering Code-Mixed Challenges: A Framework for Linguistically Driven Question Generation and Neural Based Question Answering",
   author = "Gupta, Deepak  and
     Lenka, Pabitra  and
     Ekbal, Asif  and
     Bhattacharyya, Pushpak",
   booktitle = "Proceedings of the 22nd Conference on Computational Natural Language Learning",
   month = oct,
   year = "2018",
   address = "Brussels, Belgium",
   publisher = "Association for Computational Linguistics",
   url = "https://www.aclweb.org/anthology/K18-1012",
   doi = "10.18653/v1/K18-1012",
   pages = "119--130",
     abstract = "Existing research on question answering (QA) and comprehension reading (RC) are mainly focused on the resource-rich language like English. In recent times, the rapid growth of multi-lingual web content has posed several challenges to the existing QA systems. Code-mixing is one such challenge that makes the task more complex. In this paper, we propose a linguistically motivated technique for code-mixed question generation (CMQG) and a neural network based architecture for code-mixed question answering (CMQA). For evaluation, we manually create the code-mixed questions for Hindi-English language pair. In order to show the effectiveness of our neural network based CMQA technique, we utilize two benchmark datasets, SQuAD and MMQA. Experiments show that our proposed model achieves encouraging performance on CMQG and CMQA.",
}
```