SAiDL Summer Induction Assignment

Harshvardhan Mestha

29th June, 2023

Contents

Effects of Code-Mixing on the Translation Accuracy of Large Language Models

Harshvardhan Mestha

1 Introduction

Code mixing is the process in which tokens from one language are mixed with another. For example: The phrase "ek[HI] second[EN]" is a mix of the original english phrase "one second". This is very commonly done in countries such as India and Mexico. For simpler sentences where the languages map easily, parsing is easier for models. However as the complexity of sentence, and number of code switching points increase, the model may lose accuracy. The following report details the procedures to create Code-Mixed datasets from a monolingual corpus, and a comparison between BART and mBART.

2 The training dataset

This section contains all the details pertaining to the generation of the Code-Mixed dataset. The task required me to create a code-mixed dataset from the HASOC 2019 - Hindi corpus. For creating the dataset I have used the CodeMixed Text Generator (GCM) toolkit developed by Microsoft Research. The exact procedure is detailed in the following subsections.

2.1 Getting the translations and word alignments

GCM has a web UI that uses a Flask API. To translate the sentences I used Selenium to query the instance of the web interface with a script. The translations are carried out using a Microsoft Azure Translate instance. The alignments are generated using fast_align.

Note: There is an issue with the toolkit for generating alignments, that leads to many sentences not being translated. I am yet to find a solution to this issue but I have identified the source of the problem, which the next subsection details.

2.2 Issue with the alignment generation

The alignments produced by the Azure Translator are key-value pairs of the following format -¿ (start_lang1,end_lang1)-(start_lang2,end_lang2). It contains the starting and index of the word the first language in 1 tuple and the starting and ending index of the second language in another tuple as a key-value pair. This maps every word from one language to another.

The issue arises when the translator includes a random punctuation even if the given sentence did not include one. This changes the starting and ending indexes causing an error in GCM, when it tries to convert the Azure alignments into GCM alignments.Another reason might be that the translator cannot translate the expletives present in the dataset. This shall reduce the size of the dataset considerably.

2.3 Generating the Code-Mixed Sentences

To code-mix the sentences I used Selenium to query the instance of the web interface with a script. After that we take those sentences and make a file to pass onto a transliterator as the evaluation benchmark is trained on English and Hindi-English.

A noticable problem is the fact that the translations and alignments do not work for all the sentences either further reducing the size of the dataset.

2.4 Code-Mixing Index (CMI)

I seperated the files using a script i wrote on the basis of their Code-Mixing Index which is defined as follows:

$$CMI = 100 * \frac{max(w_i)}{n - u}$$

$max(w_i)$ = number of words in the matrix language
n = total number of words, irrespective of language
u = language independent words

2.5 Transliteration

The code-mixed dataset has Devanagari characters which must be romanaised for it to work with the model used in GLUECoS. For this I used a Github repo named devanagari-to-roman-script-transliteration. The transliterated set is then loaded into GLUECoS.

3 Evaluations

The evaluation benchmark being used is the Generalized Language Evaluation Benchmark for Code-Switched NLP (GLUECoS).

The evaluation scheme shall be the MT_EN_HI benchmark.
Note: MT_EN_HI does not support BERT based models.
Note: the line "model.config.decoder_start_token_id = 2" must be added after the comment on line 397 in Code/run_seq2seq.py, else the model will not run.
The models are fine-tuned and then produce translations whose accuracy is checked against a test set. The accuracy is returned by calculating the Bilingual Evaluation Understudy Score(BLEU) score, one is the model data score, and another is run on Microsoft's GLUECoS custom evaluation script.

Models used:

1. BART (specifically facebook/bart-large)
2. mBART (specifically facebook/mbart-large-cc25)

3.1 Issues with running GLUECoS

The MT_EN_HI benchmark uses a English and Hindi-English code-mixed dataset that uses text files as the inputs for the model. Other benchmarks in the toolkit are not suitable as they work with tweet ID's for input which HASOC-2019-Hindi does not provide. However, the dataloaders for the evaluation benchmark are non functional for loading custom datasets. GLUECoS has a lot of data preprocessing done to the data from the raw dataset, which is very specific to the dataset, making it difficult to incorporate the models.

Therefore, for now I shall compare mBART v/s BART translation performance for the dataset made by Prof. Alan Black's group from CMU.

3.2 Accuracy

Model	Test Set Accuracy(BLEU%)	Evaluation Script Accuracy(BLEU%)
BART	7.48	4.62
mBART	10.38	11.01

Table 1: Accuracies for the various models.

We see that mBART performs better than BART.

4 Conclusions

To conclude, I have sucessfully produced a synthetic code-mixed dataset from the corpus provided. I have used a dataset similar to the one I have produced and observed that mBART performed better than BART. The reason for this is mBART possesses multilingual support making it much well suited to parse non-english sentence structures.

I estimate that the similar effects shall be seen with the code-mixed HASOC dataset.I also observe that the overall translation accuracy is low which can be attributed to the code-mixing, as the dataset maximises the use of Hindi-English, leading to a higher CMI. I hypothesize that the accuracy of translation models decrease with increasing the CMI, hovever I have to build better experiments to further examine this.

5 Sources/References

Rizvi, Mohd Sanad Zaki et al, "**GCM**: A Toolkit for Generating Synthetic Code-mixed Text", https://aclanthology.org/2021.eacl-demos.24

Simran Khanuja et al, "**GLUECoS**: An Evaluation Benchmark for Code-Switched **NLP**, https://www.aclweb.org/anthology/2020.acl-main.329

Ritwik Mishra, "devanagari-to-roman-script-transliteration", https://github.com/ritwikmishra/devanagari-to-roman-script-transliteration