National College of Ireland

Configuration Manual

MSc Research Project
MSc in Data Analytics

Sayok Kumar Bose
Student ID: X20187688

School of Computing
National College of Ireland

Supervisor: Mr. Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sayok Kumar Bose

Student ID: X20187688

Programme: MSc in Data Analytics **Year:** 2022

Module: MSc Research Project

Supervisor: Mr. Rejwanul Haque

Submission Due Date: 15/08/2022

Project Title: Generating Python Code from Docstrings using OpenNMT

Word Count: 6232 **Page Count** 18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sayok Kumar Bose

Date: 15/08/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Sayok Kumar Bose
Student ID: X20187688

1 Hardware Configuration

The research was conducted over Google Collab Pro sessions with high ram as the selected runtime environment. Here are the following hardware details of the underlying machine in google collab.

```
[6]  #GPU count and name
     !nvidia-smi -L

     GPU 0: Tesla P100-PCIE-16GB (UUID: GPU-651337fe-a34a-9a24-98a9-b279aacfdcfc)
```

```
     !nvidia-smi

     Mon Aug 15 09:38:36 2022
     +-----------------------------------------------------------------------------+
     | NVIDIA-SMI 460.32.03    Driver Version: 460.32.03    CUDA Version: 11.2     |
     |-------------------------------+----------------------+----------------------+
     | GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
     | Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
     |                               |                      |               MIG M. |
     |===============================+======================+======================|
     |   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
     | N/A   41C    P0    32W / 250W |    389MiB / 16280MiB |      0%      Default |
     |                               |                      |                  N/A |
     +-------------------------------+----------------------+----------------------+

     +-----------------------------------------------------------------------------+
     | Processes:                                                                  |
     |  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
     |        ID   ID                                                   Usage      |
     |=============================================================================|
     +-----------------------------------------------------------------------------+
```

```
[8]  !lscpu |grep 'Model name'

     Model name:          Intel(R) Xeon(R) CPU @ 2.20GHz
```

2 Software Configuration

For the research OpenNMT toolkit was used installation steps can be easily found here [1].

```
     !pip install --upgrade pip
     !pip install OpenNMT-py
```

[1] https://opennmt.net/OpenNMT-py/main.html#installation

3 OpenNMT

Connecting to google drive:

```
[ ] from google.colab import drive
    import os
```

```
os.environ['C /content/drive (cmd + click) = "0"
drive.mount('/content/drive', force_remount=True)
os.chdir("/content/drive/MyDrive/research-project/nmt")
```

Configuration for NMT:

config.yaml

```
## Where the samples will be written
share_vocab: true
save_data: vocab/
# Vocabulary files
src_vocab: vocab/source.pt.vocab
src_vocab_size: 50000
tgt_vocab_size: 50000
src_seq_length: 512
tgt_seq_length: 512
skip_empty_level: silent

# Training files
data:
    corpus_1:
        path_src: data/train.src.subword
        path_tgt: data/train.tgt.subword
        transforms: [filtertoolong]
    valid:
        path_src: data/valid.src.subword
        path_tgt: data/valid.tgt.subword
        transforms: [filtertoolong]


# Where to save the log file and the output models/checkpoints
log_file: train.log
save_model: model/model.code

# Stop training if it does not imporve after n validations
early_stopping: 3

# Default: 5000 - Save a model checkpoint for each n
save_checkpoint_steps: 1000

# To save space, limit checkpoints to last n
keep_checkpoint: 3

seed: 3435

# Default: 100000 - Train the model to max n steps
train_steps: 100000

# Default: 10000 - Run validation after n steps
valid_steps: 1000

# Default: 4000 - for large datasets, try up to 8000
warmup_steps: 4000
report_every: 100

# Activate TensorBoard
tensorboard: true
tensorboard_log_dir: model/tensorboard


decoder_type: transformer
encoder_type: transformer
word_vec_size: 512
rnn_size: 512
layers: 6
transformer_ff: 2048
heads: 8

accum_count: 1
optim: adam
adam_beta1: 0.9
adam_beta2: 0.998
decay_method: noam
learning_rate: 2.0
max_grad_norm: 0.0

# Tokens per batch, change if out of GPU memory
batch_size: 2048
valid_batch_size: 2048
batch_type: tokens
```

OpenNMT training and translation:

```
!onmt-main --config /content/drive/MyDrive/research-project/OpenNMT/data/config-small.yaml --model_type Transformer --auto_config train --with_eval --num_gpus 1
```

```
# Convert to a few words
# test: head -n 50 data/DataForLMG.test.subword | cut -d" " -f-15 > data/DataForLMG.test.subword.lm
cat data/DataForLMG.test.subword | cut -d" " -f-5 > data/DataForLMG.test.subword.lm

# Generation
onmt_translate -model model/model-lm_step_20000.pt -src data/DataForLMG.test.subword.lm -output data/output.subword -n_best 1 -random_sampling_topp 0.9 -beam_size 10 -gpu

# Desubwording
python3 ~/scripts/desubword.py vocab/source.model data/output.subword
```

Installing SacreBLEU for Evaluation:

```
[ ] !pip install sacrebleu
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting sacrebleu
  Downloading sacrebleu-2.2.0-py3-none-any.whl (116 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 116.6/116.6 kB 9.0 MB/s eta 0:00:00
Requirement already satisfied: lxml in /usr/local/lib/python3.7/dist-packages (from sacrebleu) (4.9.1)
Collecting colorama
  Downloading colorama-0.4.5-py2.py3-none-any.whl (16 kB)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.7/dist-packages (from sacrebleu) (
Requirement already satisfied: tabulate>=0.8.9 in /usr/local/lib/python3.7/dist-packages (from sacreble
Requirement already satisfied: regex in /usr/local/lib/python3.7/dist-packages (from sacrebleu) (2022.6
Collecting portalocker
  Downloading portalocker-2.5.1-py2.py3-none-any.whl (15 kB)
Installing collected packages: portalocker, colorama, sacrebleu
Successfully installed colorama-0.4.5 portalocker-2.5.1 sacrebleu-2.2.0
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with
```