National College of Ireland

Configuration Manual
Neural Machine Translator aided translation of the English language to under-resourced language

MSc Research Project
Data Analytics

Rajdeep Karpe
Student ID: x17164851

School of Computing
National College of Ireland

Supervisor: Prof. Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rajdeep Rajan Karpe
Student ID:	X17164851
Programme:	Data Analytics **Year:** 2019-2020
Module:	Research Project
Lecturer:	Prof. Noel Cosgrave
Submission Due Date:	12/12/2019
Project Title:	Neural Machine Translator aided translation of English to under resource language

Word Count:1196 **Page Count: 3**

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: ……

Date: 12/12/2019

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
Neural Machine Translator aided translation of the English language to under-resourced language

Rajdeep Karpe
Student ID: x17164851

Introduction

The configuration manual will guide you to build a neural machine-based translator from start. The research project has used a sequence to sequence the encoder-decoder framework[1]. For building, this translator Long Short Term Memory (LSTM) algorithm is used and the word-based translation is done from the English language to under-resourced language that is the Marathi language. The model is trained with 36,000 sentences available in the data set which is provided in an ICT solution.

The report is organized in the following way: Section 1 Will describe how to create an environment for project execution. Section 2 will provide a list of libraries used in Python and their installation command and section 3 will provide all steps performed while implementing this research project.

Section 1: Environment Setup

To implement the proposed work Graphics Processing Unit (GPU) is required. Hence, the project was implemented on Google colab, which provides 1xTesla K80 GPU, having 2496 CUDA cores with 12GB of GDDR5 VRAM for free. This helps to implement heavy codes on the cloud by using GPU with provided specifications. We need to write our code in python. Google colab provides an interface of Jupyter.

Steps to implement research project on Google Colab as follows:

1. Create new python 3 notebook.
2. Add input file first for input.
3. Start adding code cells according to process flow.
4. After entering the entire code Go to Runtime and 'Run all'.
5. Save output : File → Download .ipnby.

Section 2: Libraries used

In this section we will go through all libraries that we have used in this project and also how to install these libraries using command:

[1] https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Library Name	Command
Tensorflow	pip install –upgrade Tensorflow
Keras	pip install Keras
numpy	pip install numpy
pandas	pip install pandas
matplotlib	pip install matplotlib
sklearn	pip install sklearn
re	pip install re
nltk	pip install nltk

All of the above libraries are used with its currently updated version.

Section 3: Implementation

Dataset Collection:-

Data is collected from[2] mentioned site which provides 36,503 sentences from both the languages that are English and Marathi. We will get a zip file from a mentioned source. Source provides free access to all available files. So ethics are followed here.

Data Cleaning:-

In this process, data is cleaned to make it understandable to the machine learning algorithms. In the Data Cleaning process, we have converted all characters to lowercase so it will be easy for the machine to understand. Then we have removed all punctuation marks, special symbols, numbers, extra spaces, etc.

Tokenization:-

As our target language is Marathi, so we are going to apply start and end token on Marathi sentences. This machine will get to know about the start and the end of the sentence. This process is known as Tokenization.

Creating Dictionary:-

Here we separate all words from Marathi as well as English and then by taking each unique word from both available corpora, we sort those words alphabetically and arrange it to create a dictionary of both the languages.

Test and train split:

We have split our data into 90-10 format. 90% of data is used for training models and 10% for testing. We have used a random split of data in this research project.

Batch of Data:-

After splitting the data we need to create batches of data to pass encoder and decoder. Hence we have created batches of data with a size of 128 words per batch.

[2] http://www.manythings.org/anki/

Encoder and Decoder:-

Now we will create the encoder and decoder. The encoder will take input of English sentence and it will summarize data in its state format, and then will provide output but we don't need output we need state of summarized data as output from the encoder. That output from encoder will act as input to our decoder and will provide Marathi words for every English word as output, as our translator is working as a word-based translator. Then we will compile our model with appropriate optimizer accuracy and loss calculator. You can choose these parameters from Keras document[3]. This process of creating encoder and passing its output as an input to decoder is known as a sequence to sequence modelling for the LSTM algorithm.

Applying LSTM model on training data:

Now we will apply our LSTM model on training data. In this, the epoch decides the number of iterations. An increase in the number of epochs will provide a high learning rate for the machine which will increase the accuracy for predicting words. So basically we are training with 50 epochs first then 100 and finally 250 epochs. From this experiment, we got to know that the training of 50 epochs takes approximately 90 minutes. 100 epochs take almost 220 minutes to train but there is a significant increase in accuracy and decrease in loss which provides positive results towards our research.

Applying LSTM on test data and Validation:

After getting results on our training data we will test results on test data by using validation technique of Bilingual Evaluation Understudy (BLUE) score. The score ranges from 0 to 1.Score near to 1 tell that almost all of words are getting predicted correctly and zero means our prediction is wrong. Hence we have used BLUE score for validating result of test data.

All of the implemented python code is provided in code artefact that is in the ICT Solution[4].

[3] https://keras.io/
[4] https://github.com/RajdeepKarpe22/Research-project2