**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN CÔNG NGHỆ PHẦN MỀM**

A picture containing meter

Description automatically generated

**TÀI LIỆU HƯỚNG DẪN**

**CÀI ĐẶT VÀ BIÊN DỊCH MÃ NGUỒN**

**XÂY DỰNG MÔ HÌNH DỊCH MÁY TỪ**

**TIẾNG ANH SANG TIẾNG VIỆT**

**KHOÁ LUẬN TỐT NGHIỆP**

**GVHD:**

TS. Ngô Huy Biên

**Sinh viên thực hiện:**

1612689 – Trương Phạm Nhật Tiến

[1612689@student.hcmus.edu.vn](mailto:1612689@student.hcmus.edu.vn)

1612726 – Nguyễn Minh Trí

[1612726@student.hcmus.edu.vn](mailto:1612726@student.hcmus.edu.vn)

**Tp. Hồ Chí Minh, tháng 6 năm 2020**

**Mục lục**

[**1.** **Giới thiệu** 4](#_Toc43377908)

[**1.1** **Mục đích** 4](#_Toc43377909)

[**1.2** **Yêu cầu môi trường** 4](#_Toc43377910)

[**2.** **Cài đặt môi trường, công cụ** 5](#_Toc43377911)

[**3.** **Biên dịch mã nguồn** 6](#_Toc43377912)

**Bảng mô tả thay đổi tài liệu**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ngày** | **Phiên bản** | **Mô tả** | **Người viết** |
| 16/06/2020 | 1.0 | Hướng dẫn cách cài đặt môi trường và biên dịch mã nguồn của mô hình dịch máy | Trương Phạm Nhật Tiến  Nguyễn Minh Trí |
| 17/06/2020 | 1.1 | Cập nhật thông tin môi trường cài đặt | Trương Phạm Nhật Tiến  Nguyễn Minh Trí |

1. **Giới thiệu**
2. **Mục đích**

Tài liệu hướng dẫn cũng như cung cấp những thông tin cần thiết, chi tiết về các bước để cài đặt môi trường và biên dịch mã nguồn của mô hình dịch máy từ tiếng Anh sang tiếng Việt.

1. **Yêu cầu môi trường**

MacOS, Ubuntu hoặc Linux và cũng có thể sử dụng dịch vụ Google Colab

RAM : 25GB

1. **Cài đặt môi trường, công cụ**
2. **Chuẩn bị**

* Máy tính sử dụng hệ điều hành MacOS, Ubuntu hoặc Ubuntu

1. **Cài đặt môi trường**

* Ở trong bài này nhóm phát triển sẽ hướng dẫn cài đặt môi trường Python và các thư viện hỗ trợ trên các máy tính (sử dụng Google Colab được cung cấp sẵn nên ta có thể chuyển sang phần biên dịch mã nguồn để tiếp tục).
* Bước 1: Ta cài đặt Python 3.5 và môi trường ảo. Ta chạy 2 câu lệnh sau trên command:
  + **“sudo apt-get update”**
  + **“sudo apt install python3-dev python3-pip virtualenv”**
* Bước 2: Tạo môi trường ảo env bằng câu lệnh:
  + **“virtualenv env”**
  + Ta được kết quả có một thư mục env được tạo ra.

**A screenshot of a cell phone

Description automatically generated**

* Bước 3: Kích hoạt môi trường ảo env:
  + **“source env/bin/activate”**
  + Môi trường đã được kích hoạt có dấu hiệu như sau:

****

* Bước 4: Cài đặt các thư viện hỗ trợ:
  + **“sudo pip3 install --upgrade pip”**
  + **“sudo pip3 install --upgrade setuptools”**
  + **“sudo -H pip3 install tensorflow==1.15”**
  + **“sudo pip3 install sklearn nltk gensim matplotlib”**
* Vậy là chúng ta đã cài đặt môi trường để huấn luyện mô hình.

1. **Biên dịch mã nguồn**

* Tại đây, chúng ta sẽ tải mã nguồn mô hình từ github về máy:
  + **“git clone** [**https://github.com/nmtri1912/Model.git**](https://github.com/nmtri1912/Model.git)**”**
* **”cd Model”** **A screenshot of a cell phone

  Description automatically generated**
* Cài đặt gdown, unzip và để và giải nén word\_embedding:
  + **“pip3 install gdown”**
  + **“sudo apt-get install unzip”**
* Tiến hành tải word\_embedding:
  + **“gdown** [**https://drive.google.com/uc?id=1FbUCVL6UKDC-yyi72VUVdR4RYYaVAFQf**](https://drive.google.com/uc?id=1FbUCVL6UKDC-yyi72VUVdR4RYYaVAFQf)**”**
  + **“unzip data\_webb.zip”**

A screenshot of a cell phone

Description automatically generated

* Ta tiến hành tại 2 thư : **“vocab\_english”** và **“vocab\_vietnamese”**
  + **“mkdir vocab\_english”**
  + **“mkdir vocab\_vietnamese”**
  + Ta được thư mục Model với các thư mục con và file như sau:

**A picture containing monitor, screen, sitting, green

Description automatically generated**

* Tiến hành huấn luyện mô hình
  + **python3 train.py**

**--language\_src data/train-en-vi/train.en**

**--language\_targ data/train-en-vi/train.vi**

**--vocab\_src vocab\_english/**

**--vocab\_targ vocab\_vietnamese/**

**--word\_emb\_src word\_embedding/model\_en.bin**

**--word\_emb\_targ word\_embedding/model\_vn.bin**

**--num\_layer 2**

**--num\_hiddens 512**

**--learning\_rate 0.001**

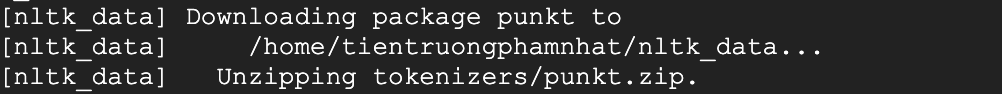
**--keep\_prob 0.85**

--beam\_width 10

--batch\_size 128

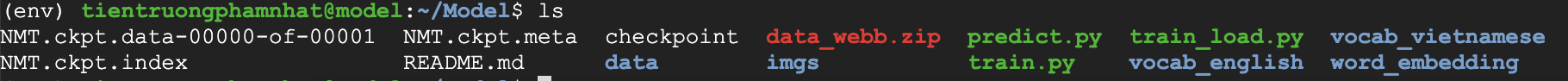
--checkpoint NMT.ckpt

* Mô hình tiến hành huấn luyện và:

A screenshot of a cell phone

Description automatically generated

* Sau khi chạy xong mỗi epoch ta sẽ có kết quả các file checkpoint như sau:



* Như vậy ta đã tiến hành xong quá trình huấn luyện mô hình dịch máy từ tiếng Anh sang tiếng Việt. Để sử dụng mô hình ta xem thêm tài liệu **“Huong\_Dan\_Trien\_Khai”**

1. **Hướng dẫn sử dụng mô hình để dịch một file từ tiếng Anh sang tiếng Việt**

* Đầu tiên ta cần tải Language Model được nhóm sinh viên phát triển để sử dụng trong lúc dịch:
  + “**gdown** [**https://drive.google.com/file/d/1-2wQJY5N1HWbFdDuznypMDfP42eWt0gA/view?usp=sharing**](https://drive.google.com/file/d/1-2wQJY5N1HWbFdDuznypMDfP42eWt0gA/view?usp=sharing&fbclid=IwAR2NhhkMdJGbecvx0_u2AE-fXVNPslxGI6l6e34MXoawX9FYlmCTXL0TyZA)
  + **pip install kenlm**
* Sau đó ta chạy đoạn lệnh:
  + **python3 predict.py**

**--language\_src data/test-2013-en-vi/tst2013.en**

**--language\_targ data/ test-2013-en-vi/tst2013.vi**

**--vocab\_src vocab\_english/**

**--vocab\_targ vocab\_vietnamese/**

**--word\_emb\_src word\_embedding/model\_en.bin**

**--word\_emb\_targ word\_embedding/model\_vn.bin**

**--num\_layer 2**

**--num\_hiddens 512**

**--learning\_rate 0.001**

**--keep\_prob 0.85**

--beam\_width 10

--batch\_size 128

--checkpoint NMT.ckpt

* Có thể ta có gặp trường hợp không đủ bộ nhớ. Khi đó ta tiến hành tách file **“tst2013.en”** thành các file nhỏ với kích thước không quá 500 dòng.