Báo cáo tiến độ tuần 9-10.

Người báo cáo: Trần Đức Thọ.

Báo cáo những thứ tìm hiểu được trong OpenNMT:

OpenNMT là gì?

OpenNMT là hệ sinh thái mã nguồn mở phục vụ cho neural machine translation và neural sequence learning.

OpenNMT hoạt động dựa trên file config nó, theo em tìm hiểu được là file có đuôi .yaml

Model hoạt động trong OpenNMT: Gồm 3 mô hình:

-RNN

-LSTM

-Transformer

Kết quả mô hình phụ thuộc vào:

-Dữ liệu train có đủ lớn không.

-Dữ liệu đem train có được sạch không.

-Số step có đủ lớn không.

Em xin phép demo với bộ data English-German:

Đầu tiên chúng ta cần phải cài bộ OpenNMT-py về máy:

pip install OpenNMT-py

Theo hướng dẫn chúng ta sử dụng 2 câu lệnh để tải bộ data về và lưu trữ:

wget https://s3.amazonaws.com/opennmt-trainingdata/toy-ende.tar.gz

tar xf toy-ende.tar.gz

wget là câu lệnh dùng để tải bộ data trên https còn tar dùng để giải nén file, sau khi giải nén chúng ta thu được 6 file

Text

Description automatically generated

Tiếp theo em khởi tạo file toy\_en\_de.yaml để config:

*# toy\_en\_de.yaml  
  
## Where the samples will be written*save\_data: toy-ende/run/example  
*## Where the vocab(s) will be written*src\_vocab: toy-ende/run/example.vocab.src  
tgt\_vocab: toy-ende/run/example.vocab.tgt  
*# Prevent overwriting existing files in the folder*overwrite: False  
  
*# Corpus opts:*data:  
 corpus\_1:  
 path\_src: toy-ende/src-train.txt  
 path\_tgt: toy-ende/tgt-train.txt  
 valid:  
 path\_src: toy-ende/src-val.txt  
 path\_tgt: toy-ende/tgt-val.txt  
*# Vocabulary files that were just created*src\_vocab: toy-ende/run/example.vocab.src  
tgt\_vocab: toy-ende/run/example.vocab.tgt  
  
  
  
*# Where to save the checkpoints*save\_model: toy-ende/run/model  
save\_checkpoint\_steps: 50  
train\_steps: 100  
valid\_steps: 50

data: là nơi em lưu đường dẫn file train và val do máy yếu em chỉ để 100 step train và 50 step val:

tiếp theo em xây dựng vocabulary cho mô hình bằng dòng lệnh:

onmt\_build\_vocab -config toy\_en\_de.yaml -n\_sample 10000

-n\_sample 10000 ở đây nghĩa là em chỉ dùng 10000 câu để xây dựng vocabulary

Bước tiếp theo em đưa mô hình vào train bằng lệnh

onmt\_train -config toy\_en\_de.yaml

quá trình này sẽ tốn khá nhiều thời gian

sau khi chạy xong em bắt đầu cho dịch mô hình:

onmt\_translate -model toy-ende / run / model\_step\_100.pt -src toy-ende / src-test.txt -output toy-ende / pred\_100.txt -verbose

do mô hình đào tạo không sâu ( chỉ 10000 chữ,100 step train,50 stap val) nên kết quả thu được rất tồi, hầu như không dịch được

