By Vivien BOGNE TIENOUE
EPITA, DSA
January 30, 2021

Introduction

In this session, we propose to create a very first model for Neural Machine Translation

I. Tokenization

Exercise 1:

Q1: Explain in a few words the goal and the algorithm of BPE.

- The Goal : *sub-word tokenization; in fact, "Byte-Pair Encoding (BPE) is an unsupervised sub-word tokenization technique, commonly used in neural machine translation and other NLP tasks."*[1] [2] *Words in Languages are built up of different sub-words and taking those sub-words enhance the quality of embeddings on a NMT.*
- The Algorithm: *It is a data compression algorithm; it "iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte."*[3]

Q2: Prepare 3 BPE models, associated with 3 code sizes: 16k, 32k and 48k. What are the exact amount of code done for each one?

After prepare the model we have :

```
! ls ./codes/ -ltr
```

```
total 1032
-rw-r--r-- 1 root root 160388 Jan 31 14:34 codes_files_16k.bpe
-rw-r--r-- 1 root root 345768 Jan 31 14:36 codes_file_32k.bpe
-rw-r--r-- 1 root root 543107 Jan 31 14:38 codes_file_48k.bpe
```

- Amount of code for 16k : 16001 (16k + one line for the version)

```
15998   alk aline</w>
15999   al ?</w>
16000   affin ité</w>
16001   ab ond
16002
```

- Amount of code for 32k: 32001 (32k + one line for the version)

```
31999   pover ty.</w>
32000   pous soir</w>
32001   poursuiv ra</w>
32002
```

[1] *Investigating the Effectiveness of BPE: The Power of Shorter Sequences*, Matthias Galle
[2] See Annex figure 1 for the illustration.
[3] *Neural Machine Translation of Rare Words with Subword Units*, Rico Sennrich and Barry Haddow and Alexandra Birch School of Informatics, University of Edinburgh

- Amount of code for 48k : 48001 (48k + one line for the version)

```
47998   moder ation</w>
47999   mobil is
48000   mo te</w>
48001   mo s</w>
48002
```

Since I used the syntax :

```
# First 16k BPE model
! subword-nmt learn-joint-bpe-and-vocab --input ./drive/MyDrive/train.fr ./drive/MyDrive/train.en -s 16000 -o ./codes/codes_files_16k.bpe  --write-vocabular
# First 32k BPE model
! subword-nmt learn-joint-bpe-and-vocab --input ./drive/MyDrive/train.fr ./drive/MyDrive/train.en -s 32000 -o ./codes/codes_file_32k.bpe  --write-vocabulary
# Third 48k BPE model
! subword-nmt learn-joint-bpe-and-vocab --input ./drive/MyDrive/train.fr ./drive/MyDrive/train.en -s 48000 -o ./codes/codes_file_48k.bpe  --write-vocabulary
```

The vocabs file has be created with the number of occurrences behind each subwords. For example

```
1 the 50039
2 of 36050
3 and 23687
```

Q3: Open the code files compare then, what can you conclude?

After opening the different codes_files, the observation is clear: codes_files_16k is included in codes_files_32k which is also included in code_files_48k. Then the more is the number of iterations on BPE, the more the sub-word tokenizer will be.

Q4: Generate 3 datasets (i.e. train, dev, test and vocabularies) associated to the 3 models, compare the amount of data.

To generate we apply our different models to the train, dev, test and we get the vocab of train.

For Each model we have :

- 2 train : fr and en.
- 2 dev : fr and en
- 2 test : fr and en
- 2 vocab: fr and en.

The model with the size 48k have the biggest corresponding files and the smallest on the 16k.

```
[17] ! ls ./transformer_model_16k/ -ltr

total 18424
-rw-r--r-- 1 root root  138822 Jan 31 14:35 vocab_16k.fr
-rw-r--r-- 1 root root  120557 Jan 31 14:35 vocab_16k.en
-rw-r--r-- 1 root root 9583101 Jan 31 14:39 train.BPE.fr
-rw-r--r-- 1 root root 8232181 Jan 31 14:39 train.BPE.en
-rw-r--r-- 1 root root  191255 Jan 31 14:39 dev.BPE.fr
-rw-r--r-- 1 root root  166582 Jan 31 14:39 dev.BPE.en
-rw-r--r-- 1 root root  187020 Jan 31 14:39 test.BPE.fr
-rw-r--r-- 1 root root  160269 Jan 31 14:39 test.BPE.en
-rw-r--r-- 1 root root   38676 Jan 31 14:39 vocab.fr
-rw-r--r-- 1 root root   31494 Jan 31 14:39 vocab.en
```

```
! ls ./transformer_model_32k/ -ltr

total 19668
-rw-r--r-- 1 root root   263188 Jan 31 14:36 vocab_32k.fr
-rw-r--r-- 1 root root   223524 Jan 31 14:36 vocab_32k.en
-rw-r--r-- 1 root root 10145061 Jan 31 14:42 train.BPE.fr
-rw-r--r-- 1 root root  8685853 Jan 31 14:42 train.BPE.en
-rw-r--r-- 1 root root   203273 Jan 31 14:42 dev.BPE.fr
-rw-r--r-- 1 root root   176695 Jan 31 14:42 dev.BPE.en
-rw-r--r-- 1 root root   199050 Jan 31 14:42 test.BPE.fr
-rw-r--r-- 1 root root   169707 Jan 31 14:42 test.BPE.en
-rw-r--r-- 1 root root    28948 Jan 31 14:42 vocab.fr
-rw-r--r-- 1 root root    23957 Jan 31 14:43 vocab.en
```

```
[19] ! ls ./transformer_model_48k/ -ltr

total 20680
-rw-r--r-- 1 root root   378642 Jan 31 14:38 vocab_48k.fr
-rw-r--r-- 1 root root   317135 Jan 31 14:38 vocab_48k.en
-rw-r--r-- 1 root root 10564341 Jan 31 14:43 train.BPE.fr
-rw-r--r-- 1 root root  9061123 Jan 31 14:43 train.BPE.en
-rw-r--r-- 1 root root   212447 Jan 31 14:43 dev.BPE.fr
-rw-r--r-- 1 root root   184918 Jan 31 14:43 dev.BPE.en
-rw-r--r-- 1 root root   207372 Jan 31 14:43 test.BPE.fr
-rw-r--r-- 1 root root   177339 Jan 31 14:43 test.BPE.en
-rw-r--r-- 1 root root    24989 Jan 31 14:43 vocab.fr
-rw-r--r-- 1 root root    21131 Jan 31 14:43 vocab.en
```

II. Data file configuration:

This is the content of my data.yml for the model_32k:

```yaml
model_dir: transformer_model_32k/run/

data:
  train_features_file: transformer_model_32k/train.BPE.fr
  train_labels_file: transformer_model_32k/train.BPE.en
  eval_features_file: /content/drive/MyDrive/data_translation/ep7_dev.BPE.fr
  eval_labels_file: /content/drive/MyDrive/data_translation/ep7_dev.BPE.en
  source_vocabulary: transformer_model_32k/vocab.fr
  target_vocabulary: transformer_model_32k/vocab.en

train:
  batch_size: 1024
  save_checkpoints_steps: 1000
  maximum_features_length: 50
  maximum_labels_length: 50

eval:
    eval_delay: 3600  # Every 1 hour
    external_evaluators: BLEU
    export_on_best: BLEU
infer:
    batch_size: 32
```

It is the same for the 2 other models only the path location of the file is changing according to the model. And the command used to train the model is :

```
# Training
! onmt-main --model_type Transformer --config ./drive/MyDrive/data.yml --auto_config train --with_eval --num_gpus 1
```

```
# Training 32k
! onmt-main --model_type Transformer --config ./drive/MyDrive/data_32k.yml --auto_config train --with_eval --num_gpus 1
```

```
# Training 48k
! onmt-main --model_type Transformer --config ./drive/MyDrive/data_48k.yml --auto_config train --with_eval --num_gpus 1
```

I had to save the model every time on my local drive to avoid loosing the training process. Indeed after 12 hours google colab stop the session. To continue the training model I just added the number of step to the files data.yml then train again.

III. Translate the model obtained.

```
# inference (Translation)
! onmt-main --config ./drive/MyDrive/data.yml --auto_config infer --features_file ./transformer_model_16k/test.BPE.fr --predictions_file /content/transformer_model_16k/test_pred.BPE.en
```

```
# inference (Translation)
! onmt-main --config ./drive/MyDrive/data_32k.yml --auto_config infer --features_file ./transformer_model_32k/test.BPE.fr --predictions_file /content/transformer_model_32k/test_pred.BPE.en
```

```
# inference (Translation)
! onmt-main --config ./drive/MyDrive/data_48k.yml --auto_config infer --features_file ./transformer_model_48k/test.BPE.fr --predictions_file /content/transformer_model_48k/test_pred.BPE.en
```

V. Detokenization

```
# detokenization
cat ./transformer_model_16k/test_pred.BPE.en | sed "s/@@ //g" > ./transformer_model_16k/test.en
cat ./transformer_model_32k/test_pred.BPE.en | sed "s/@@ //g" > ./transformer_model_32k/test.en
cat ./transformer_model_48k/test_pred.BPE.en | sed "s/@@ //g" > ./transformer_model_48k/test.en
```

VI. Evaluation Using BLEU

Using the script multi-bleu.pl

	Model 16k	Model 32k	Model 48k
BLEU Score %	27.8	27.3	27.1

So the model BPE 48k is the one with the smallest Bleu Score

Exercice 2:

Q1: translate and evaluate the EMEA_tst file. What is the BLEU score?:

In the Evaluation above we used the ep7_test. Now using EMEA_tst file we have :

	Model 16k	Model 32k	Model 48k
BLEU Score %	30.1	30.3	30.5

For example this is how we compute the evaluation for the this model_16k

```
# Evaluation Using BLEU 16k
! perl ./multi-bleu.perl /content/transformer_model_16k/train.en.bpe < ./transformer_model_16k/test.en
```

```
BLEU = 30.11, 8.2/0.3/0.1/0.0 (BP=0.395, ratio=0.518, hyp_len=19499, ref_len=37618)
It is not advisable to publish scores from multi-bleu.perl.  The scores depend on your tokenizer, which is unlikely to be repro
```

Q2: re-launch the training process by using EMEA_dev file as valid file and re-translate and re-evaluate the EMEA_tst and ep7_tst files. Is there a strong difference between the two experiments?

First we used Ep7 as validation file.

Now we the use of EMEA_dev (Changing the input validation on data.yml).

Using only the model 16k we have the following results :

BLEU Score %	Evaluation of EP7_tst	Evaluation of EMEA_tst
Valid : EP7_dev	27.6	30.6
Valid = EMEA_dev	21.3	20.2

No the difference is not that strong.

Q3: Re-evaluate all the translations using other metrics often used for Machine Translation. Chose two other metrics, score the files, explain why you chose them and explain them.

First we are using the script mteval-v14c.pl

Then we will choose AMBER (Modified BLEU, Enhanced Ranking Metric), which is based on the metric BLEU but incorporates recall, extra penalties, and some text processing variants.

AMBER %	Evaluation of EP7_tst	Evaluation of EMEA_tst
Valid = EP7_dev	30.6	30.2
Valid = EMEA_dev	26.8	22.9

Finally we will choose METEOR (Metric for Evaluation of Translation with Explicit Ordering), which is by far the most widely employed human-targeted metric in MT. I

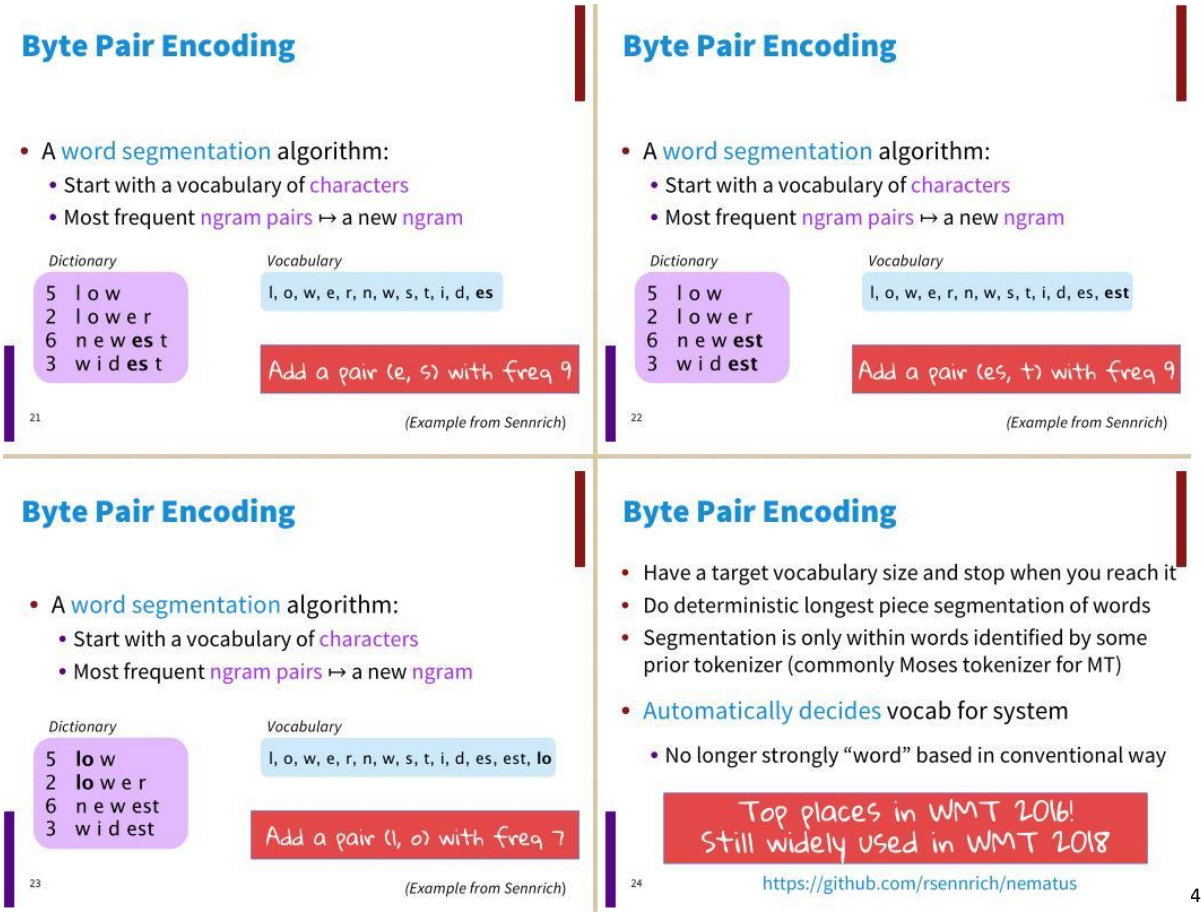
METEOR Score %	Evaluation of EP7_tst	Evaluation of EMEA_tst
Valid = EP7_dev	29.5	30.9
Valid = EMEA_dev	27.6	26.7

References:

1. The script : https://colab.research.google.com/drive/1YFODjXY06FhmhJguHKdbTbQ2_0tzVTWr?authuser=3#scrollTo=9ceKWm01Avbq

Annex:

Figure 1: The foundation of NLP-Subword Model



[4]

[4] Source : https://www.programmersought.com/article/47414443821/