Progress Update

Yu-Hsiang Lin
Language Technologies Institute
Carnegie Mellon University
yuhsianl@andrew.cmu.edu

Abstract

Place holder [**?**].

1 TTR using Language-independent tokenization

Tokenizer: SentencePiece (https://github.com/google/sentencepiece)

Data: Spanish–English (/home/gneubig/exp/transfer-exp/data/spa_eng)

For debugging purposes, we first consider English, and use only the training set (ted-train.mtok.eng) to train the tokenizer model. The number of sentences in this training set is 196,036.

Model: The model type is unigram, vocabulary size is 8,000, character coverage is 100%.

If we run tokenizer on the same training dataset using this model, the TTR result is:

distinct token count = 7,989

token count = 5,386,583

TTR = 0.001483

Note that a few sample high-/low-frequency tokens are shown in Figures 1 and 2. Actually when it performs the tokenization, the UNK, BOS, EOS have been removed (Figure 3).

Should probably further remove things like “quot”, “_&”, and punctuations.

If I increase the vocabulary size to 16,000, the result is:

distinct token count = 15,981

token count = 5,232,195

TTR = 0.003054

Give more example of vocabularies in the middle in Figure 4.

(It seems that increasing vocabulary size of model will increase TTR, which indicates that the previous vocabulary size may be too small to be representative.)

vocabulary size = 32,000

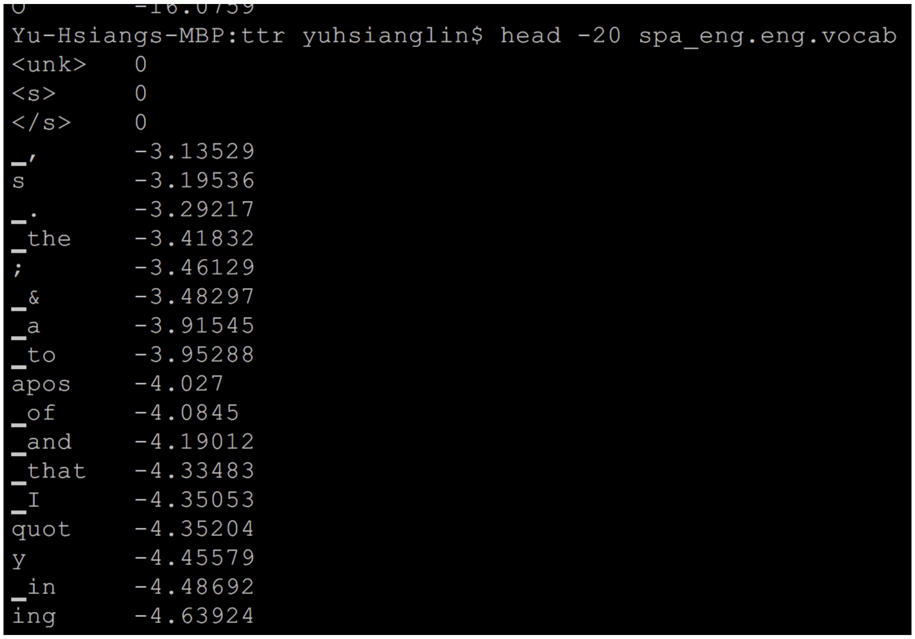
distinct token count = 29,765

token count = 5,177,707

TTR = 0.005749

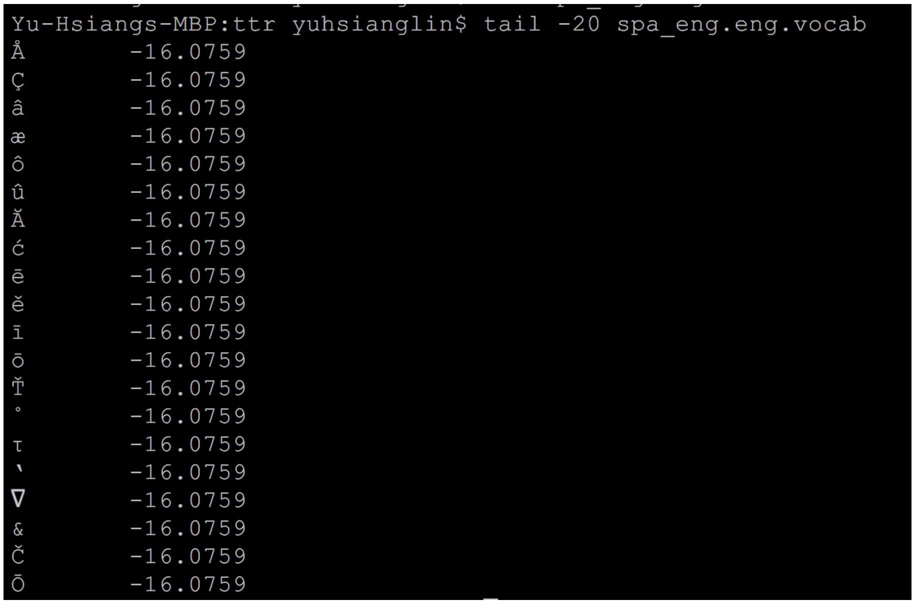
Vocabulary size larger than 32,000 seems to be too large. Encounter error during training: the training ends early and reports error

RuntimeError: Internal: /Users/travis/build/google/sentencepiece/src/trainer_interface.cc(343) [(trainer_spec_.vocab_size()) == (model_proto->pieces_size())]



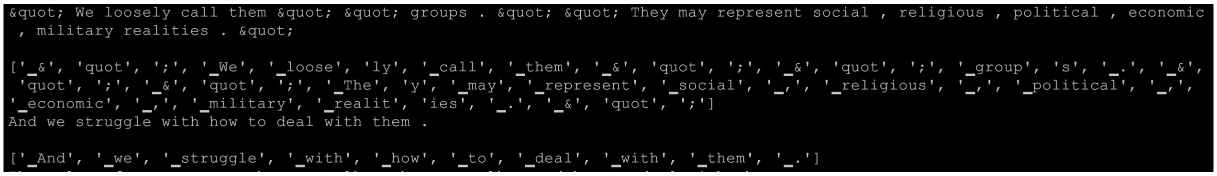
```
Õ       -16.0759
Yu-Hsiangs-MBP:ttr yuhsianglin$ head -20 spa_eng.eng.vocab
<unk>   0
<s>     0
</s>    0
▁,      -3.13529
s       -3.19536
▁.      -3.29217
▁the    -3.41832
;       -3.46129
▁&      -3.48297
▁a      -3.91545
▁to     -3.95288
apos    -4.027
▁of     -4.0845
▁and    -4.19012
▁that   -4.33483
▁I      -4.35053
quot    -4.35204
y       -4.45579
▁in     -4.48692
ing     -4.63924
```

Figure 1: Top 20.



```
Yu-Hsiangs-MBP:ttr yuhsianglin$ tail -20 spa_eng.eng.vocab
Å       -16.0759
Ç       -16.0759
â       -16.0759
æ       -16.0759
ô       -16.0759
û       -16.0759
Ă       -16.0759
ć       -16.0759
ē       -16.0759
ě       -16.0759
ī       -16.0759
ō       -16.0759
Ť       -16.0759
°       -16.0759
τ       -16.0759
‛       -16.0759
∇       -16.0759
&       -16.0759
Č       -16.0759
Õ       -16.0759
```

Figure 2: Tail 20.

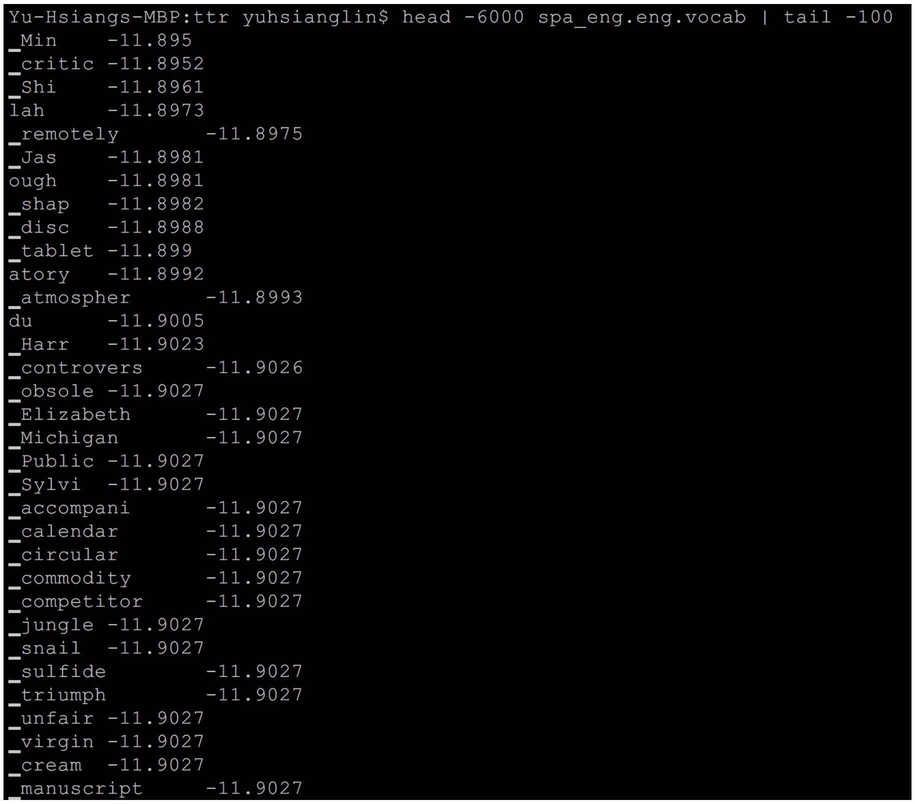


```
&quot; We loosely call them &quot; &quot; groups . &quot; &quot; They may represent social , religious , political , economic
 , military realities . &quot;

['▁&', 'quot', ';', '▁We', '▁loose', 'ly', '▁call', '▁them', '▁&', 'quot', ';', '▁&', 'quot', ';', '▁group', 's', '▁.', '▁&',
 'quot', ';', '▁&', 'quot', ';', '▁The', 'y', '▁may', '▁represent', '▁social', '▁,', '▁religious', '▁,', '▁political', '▁,',
'▁economic', '▁,', '▁military', '▁realit', 'ies', '▁.', '▁&', 'quot', ';']
And we struggle with how to deal with them .

['▁And', '▁we', '▁struggle', '▁with', '▁how', '▁to', '▁deal', '▁with', '▁them', '▁.']
```

Figure 3: Example.



```
Yu-Hsiangs-MBP:ttr yuhsianglin$ head -6000 spa_eng.eng.vocab | tail -100
▁Min    -11.895
▁critic -11.8952
▁Shi    -11.8961
lah     -11.8973
▁remotely       -11.8975
▁Jas    -11.8981
ough    -11.8981
▁shap   -11.8982
▁disc   -11.8988
▁tablet -11.899
atory   -11.8992
▁atmospher      -11.8993
du      -11.9005
▁Harr   -11.9023
▁controvers     -11.9026
▁obsole -11.9027
▁Elizabeth      -11.9027
▁Michigan       -11.9027
▁Public -11.9027
▁Sylvi  -11.9027
▁accompani      -11.9027
▁calendar       -11.9027
▁circular       -11.9027
▁commodity      -11.9027
▁competitor     -11.9027
▁jungle -11.9027
▁snail  -11.9027
▁sulfide        -11.9027
▁triumph        -11.9027
▁unfair -11.9027
▁virgin -11.9027
▁cream  -11.9027
▁manuscript     -11.9027
```

Figure 4: Middle (6000+).