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())]

```
Ö       -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
ĕ       -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
Ō       -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+).

2 Ranking problem

[Query]

Task language: the (low-resource) source language that we want to translate from. (In our current dataset, the target language is always English)

[Documents]

Candidate auxiliary languages: the (high-resource) source language whose data we are going to add to that of the task language to (1) increase the training data size and/or (2) to regularize overfitting to small task language data.

[Features]

Query features:

1. Task language dataset size
2. Task language TTR

Document features:

1. Auxiliary language dataset size
2. Auxiliary language TTR

Query-document interaction features:

1. Word-level overlap between task and auxiliary language datasets
2. Character-level overlap between task and auxiliary language datasets
3. URIEL distance between task and auxiliary language (dataset independent)
4. More...

[Rank]

The ranking of the auxiliary languages according to the BLEU scores obtained by training a model using (task language dataset + auxiliary language dataset). Note that the **higher** the rank the better the auxiliary language is.

[Query group]

The data associated with a query consist of multiple data points coming from all the candidate auxiliary languages. Each data point has features of a auxiliary language and its interactions with the task language that defines this query, as well as the rank of this auxiliary language. Such a set of data points is called a query group.