byte_pair_embedding-2

July 12, 2021

The words with “most similar words” that are understandable are natutuhan and bayanihan. Natutuhan word have list of encodings are ”_natu“, and”tuhan” and it’s most similar words that make sense are ”_matu“,”_natutu“,”tunan“,”klasan (tu-klasan)“, and”tukoy“. Bayanihan on the other hand have encodings”_bayan“, and”ihan” and most similar words that make sense “lalawigan”, “bayan”, “pilipinas”, “pagawa”, and “senso (asenso)”.

The words with “most similiar words” that does not make sense are tau-tauhan, nagloloko, and talahanayan. The following are their word encodings: 1. Tau-tauhan: a) _ta b) u c) - d) ta e) uhan 2. Nagloloko: a) _nag b) lo c) lo d) ko 3. Talahanayan: a) _tala b) ha c) nayan

The most similar words for 1 and 2 are in at most 4 characters.

In these words, it is important how does the bpemb captures the encodings of the word. For words like bayanihan and natutuhan, it returns a similar words that are understandable because their encoding is close to what is expected.

For words like tau-tauhan, nagloloko, and talahanayan, it returns most similar words that are incomprehensible or far from the meaning. I think it has something to do with the corpus. The token lo have more counts than the word loko and the word nayan have more counts than the word hanay. For the word tau-tauhan, it does not capture the word tao because of the difference in a single letter o->u.

Thus, if I am to create an MP with NLP, I have to consider what context should I consider or what corpus should I use.

[ ]:

```
!pip install bpemb
```

[10]:

```
from bpemb import BPEmb
bpemb_tl = BPEmb(lang="tl", vs=10000, dim=300)
```

[37]:

```
def print_list(list_var):
  for element in list_var:
    print(element)
```

[39]:

```
# Meaning - "nagkaroon ng kaalaman sa isang bagay o kasanayan sa isang sining, hanapbuhay, at iba pa sa pamamagitan ng pag-aaral, pagbabasá, at katulad na karanasan" - https://www.tagaloglang.com/natutuhan/
# salitang-ugat - matuto
encoding = bpemb_tl.encode("natutuhan")
print_list(encoding)
```

```python
print(" ")
print_list(bpemb_tl.most_similar(encoding))
```

```
 natu
tuhan

(' matu', 0.5096007585525513)
(' natutu', 0.4488203525543213)
('tunan', 0.4432719945907593)
('klasan', 0.38304567337036133)
(' tinu', 0.3500446081161499)
('tukoy', 0.33783847093582153)
(' komple', 0.3288344442844391)
(' kaagad', 0.3076633810997009)
('hayan', 0.30554503202438354)
('lungan', 0.29603707790374756)
```

[40]:
```python
# In english puppet - manikin; figurehead - https://www.tagalog.com/words/tau-tauhan.php
# salitang-ugat - tao
encoding = bpemb_tl.encode("tau-tauhan")
print_list(encoding)
print(" ")
print_list(bpemb_tl.most_similar(encoding))
```

```
 ta
u
-
ta
uhan

('ung', 0.3598358631134033)
('i', 0.3571832776069641)
(' pag', 0.34716206789016724)
('tag', 0.34570395946502686)
('ba', 0.343934893608093 26)
('ku', 0.3416160047054291)
('a', 0.339763343334198)
('lang', 0.32915276288986206)
(' u', 0.32822215175628662)
('un', 0.3269190192222595)
```

[41]:
```python
# In english - "fooling around" - Google translate
# salitang-ugat - loko
encoding = bpemb_tl.encode("nagloloko")
print_list(encoding)
print(" ")
```

```
print_list(bpemb_tl.most_similar(encoding))
```

```
 nag
lo
lo
ko

('ri', 0.3619576692581177)
('ro', 0.3606819212436676)
(' ay', 0.34522873163223267)
(' at', 0.33847755193710327)
('u', 0.3334447741508484)
(' pag', 0.3233245611190796)
(' na', 0.32325172424316406)
('la', 0.3192806839942932)
('sa', 0.3156728744506836)
(' ba', 0.31331107020378113)
```

[42]:

```
# In english - "nagtutulungan bilang isang pamayanan upang makamit ang isang␣
 ↪karaniwang layunin" - https://www.shopcambio.co/blogs/news/
 ↪the-bayanihan-spirit-7-filipino-social-enterprises-changing-communities#:~:
 ↪text=Bayanihan%20(buy%2Duh%2Dnee,to%20achieve%20a%20common%20goal.
# salitang-ugat - bayan
encoding = bpemb_tl.encode("bayanihan")
print_list(encoding)
print(" ")
print_list(bpemb_tl.most_similar(encoding))
```

```
 bayan
ihan

(' klaseng', 0.5418272018432617)
(' lalawigan', 0.48942610621452332)
(' kawan', 0.43971776962280273)
(' senso', 0.40939927101135254)
(' lanao', 0.39921247959136963)
('bayan', 0.3846386671066284)
(' pilipinas', 0.3784087300300598)
('yan', 0.3699503540992737)
(' pagawa', 0.36590802669525146)
(' norte', 0.34622740745544434)
```

[45]:

```
# Meaning - ginagamit sa pagtala ng datos. Halimbawa, Google Sheets Excel
# salitang-ugat - hanay
encoding = bpemb_tl.encode("talahanayan")
print_list(encoding)
print(" ")
```

```
print_list(bpemb_tl.most_similar(encoding))
```

```
 tala
ha
nayan

(' kaha', 0.38530465960502625)
(' patu', 0.3634767532348633)
('mak', 0.3576555848121643)
('tala', 0.3431604504585266)
('nay', 0.3347127437591553)
('nayang', 0.32372236251831055)
('mpas', 0.3170166015625)
('halagahan', 0.3015357255935669)
(' sumusunod', 0.29492130875587463)
('tang', 0.28682249784469604)
```