mp1_BPE

July 11, 2021

1 MP1: BYTE PAIR ENCODING

1.1 Submitted by: Juachon, Jean Philip L. (12083496)

[1]:
```
#import the library
from bpemb import BPEmb
bpemb_tl = BPEmb(lang = "tl", vs = 10000, dim = 300) #set vocabulary size to␣
↪10,000 and dimensions to 300
```

[2]:
```
bpemb_tl.emb
```

[2]:
```
<gensim.models.keyedvectors.KeyedVectors at 0x7f81586e9ed0>
```

[3]:
```
bpemb_tl.most_similar(bpemb_tl.encode("kagandahan"))
```

[3]:
```
[(' patimpalak', 0.6610562205314636),
 ('reyna', 0.5789141654968262),
 (' nanalo', 0.40153196454048157),
 (' kilalang', 0.3779382109642029),
 (' anghel', 0.3223453760147095),
 (' kalusugan', 0.3222472369670868),
 (' pilipinas', 0.29742032289505005),
 (' karangalan', 0.2955053448677063),
 (' likas', 0.29104721546173096),
 (' host', 0.2897968590259552)]
```

[4]:
```
bpemb_tl.most_similar(bpemb_tl.encode("kasintahan"))
```

[4]:
```
[(' sanda', 0.42032358050346375),
 ('lingan', 0.38719937205314636),
 (' kaibigan', 0.3487645983695984),
 ('mbi', 0.3459661602973938),
 ('pun', 0.3354654610157013),
 ('unga', 0.3213435411453247),
 (' limi', 0.3149784505367279),
 (' asawa', 0.3118269145488739),
 ('lili', 0.2897605001926422),
 (' pun', 0.2752463221549988)]
```

[5]:
```
bpemb_tl.most_similar(bpemb_tl.encode("pinagprituhan"))
```

[5]:
```
[('aaralan', 0.38834348320961),
 (' mapag', 0.36394619941711426),
 ('kunan', 0.3286280930042267),
 (' napag', 0.30512624979019165),
 ('sanib', 0.304129034280777),
 ('kukunan', 0.28864216804504395),
 ('kuran', 0.28255900740623474),
 ('pasya', 0.27607423067092896),
 ('lunan', 0.27404335141181946),
 ('lagay', 0.2674868106842041)]
```

[6]:
```
bpemb_tl.most_similar(bpemb_tl.encode("kaluwalhatian"))
```

[6]:
```
[(' malu', 0.3785698115825653),
 ('pirit', 0.3728918433189392),
 (' paglu', 0.337922066450119),
 (' pagda', 0.3334289491176605),
 ('lu', 0.3295641839504242),
 (' lu', 0.3187568783760071),
 (' espiritu', 0.3052479326725006),
 ('ulu', 0.2874179780483246),
 (' diyos', 0.28604450821876526),
 (' ka', 0.27606600522994995)]
```

[7]:
```
bpemb_tl.most_similar(bpemb_tl.encode("kapayapaan"))
```

[7]:
```
[(' kasunduang', 0.4246227741241455),
 (' kaayusan', 0.42337659001350403),
 (' pagpapanatili', 0.42105239629745483),
 ('yapaan', 0.4030686020851135),
 (' kasunduan', 0.4020146429538727),
 (' katarungan', 0.40071964263916016),
 (' digmaan', 0.3999422788619995),
 (' pagkakaisa', 0.3641497492790222),
 (' kaligtasan', 0.3524089455604553),
 ('natili', 0.34908100962638855)]
```

2 Conclusion

Based on the results above for the following words: a) kagandahan b) kasintahan c) pinagprituhan d) kaluwalhatian e) kapayapaan There are words that have results with meaningful similarities while others do not. However, it is good enough for most cases. For example, for the first word “kagandahan”, it was able to detect similar words with regards or is related to beauty such as “reyna”(queen), “anghel”(angel), and even the location “pilipinas”(Philippines), for the second word “kasintahan”, it was also able to find the similar words with regards to relationship such as

“kaibigan”(friend), “asawa”(spouse), however, it was also found subwords with regards to location such as “pun”(punta-han; location), verbs such as “mbi-tahan”(invite). Other than that, it was also able to relate the third word “pinagprituhan” to a location, with “aaralan” (place for studying) as it’s most similar word. It was also successful in identifying similar words with regards to “luwalhati” which are “espiritu”, and “diyos”. Lastly, for the word “kapayapaan”, it was also able to able to find words with regards to peace, war, and understanding or agreement. The algorithm was also able to detect and differentiate the affixes “ka-”, “-han”, and “-an” which usually refers to a state, location, and verb. Furthermore, there are also subwords that are totally not related to the input words. This happens probably because of the vocabulary size which is 10,000 only, as the vocabulary size equates to the sum of the number of BPE merge operations and the number of characters in the training data. Thus, if the vocabulary size is small, then the number of the merge operations will also be small, with this logic, a small vocabulary size will produce a smaller n-gram (e.g. unigram, bigram, etc.) while a bigger vocabulary size will create more merge operations thus, resulting into better encoding/representation of the words especially if the base form of the word is long.

Reference(s): Heinzerling, B., & Strube, M. (n.d.). BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. LREC Conferences. Retrieved July 11, 2021, from https://www.lrec-conf.org/proceedings/lrec2018/pdf/1049.pdf