Machine Translation

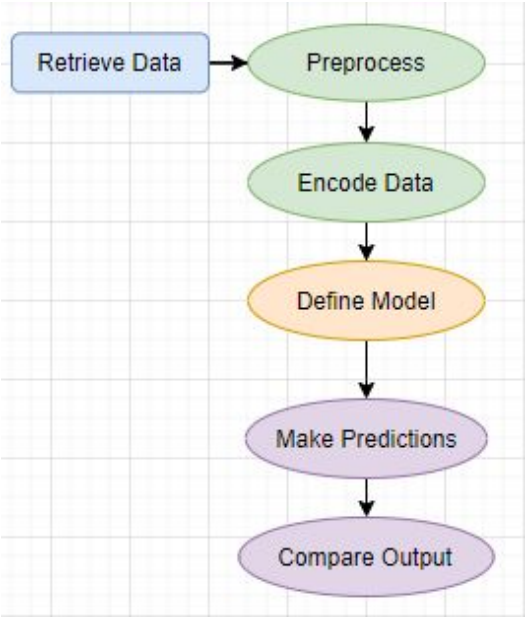
By: Sunny Bhandal

Introduction

Develop a machine translation model that will translate French to English.

Use Case:

- Educational purposes for language courses where students would fix the machines mistakes.



The Data

Data Structure

```
Go.	Va !	CC-BY 2.0 (France) Attribution: tatoeba.org #2877272 (CM) & #1158250 (Wittydev)
Go.	Marche.	CC-BY 2.0 (France) Attribution: tatoeba.org #2877272 (CM) & #8090732 (Micsmithel)
Go.	Bouge !	CC-BY 2.0 (France) Attribution: tatoeba.org #2877272 (CM) & #9022935 (Micsmithel)
```

Preprocessing

- Remove Punctuation
- Lowercase sentences
- Keep sentences less than 6 words
- Use Tokenizer to obtain vocabulary size
- Encode and pad sequences for modelling

After Preprocessing

```
array([['go', 'va '],
       ['go', 'marche'],
       ['go', 'bouge '],
       ...,
       ['no one wants to fight', 'personne ne veut se battre'],
       ['no one wants to speak', 'personne ne veut parler'],
       ['no one was helping us', 'personne ne nous aidait']],
      dtype='<U349')
```

Modelling/Predicting

LSTM Model (Embedding -> LSTM -> RepeatVector -> LSTM -> Dense Layer)

- 20% validation split
- 10 Epochs
- Batch Size: 512

Predicting

- Initially obtained numerical predictions of the vocabulary
- Convert numerical arrays into corresponding words

Example Output

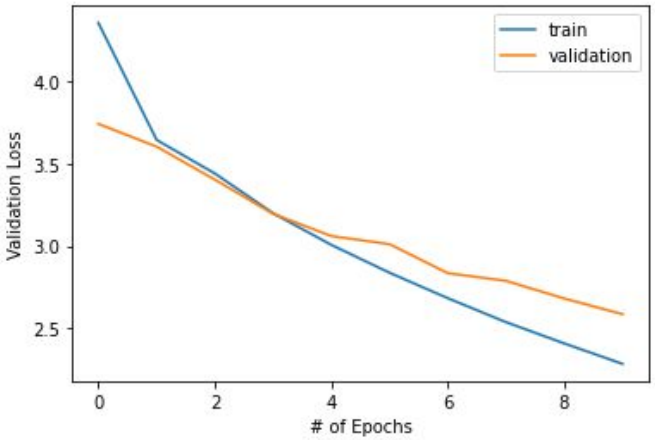
```
array([[ 1, 19,  7,  0,  0,  0],
       [ 1,  1,  2,  0,  0,  0],
       [ 2, 23,  4,  4,  0,  0],
       ...,
       [ 1, 48,  2,  2,  0,  0],
       [17, 16,  0,  0,  0,  0],
       [ 1, 19,  4, 22,  0,  0]])
```

0 = “ “

1 = “The”

Results

Model Loss



Sample Output

	actual	predicted
8473	i dont fear death	im not no
1288	where are my bags	wheres your
7642	we had a bad day	we have a
6009	help me tom	is tom
2664	get ready for a shock	i like my
4647	were satisfied	were are
5732	peel the orange	the
1556	dont open your book	dont have my
6105	i wont be late	im not so
7056	has tom taught french	is tom

Challenges

Obtaining predictions, initially my predictions were only 0’s.

- Not using enough data

Obtaining meaningful predictions.

- Words had multiple translations (EX: Go = Va, March, Bouge)
- Stopwords such as “the” may have skewed the results.
- Underfit model (less than optimal epochs used)

(Stretch) Twitter Translation

Translate tweets from English to Spanish using the Google Translate package.

After preprocessing

	clean_tweet	text
0	the iraqi side was hella peaceful	the iraqi side was hella peaceful :)
1	Make it happen then	@CbassCOD Make it happen then :)
2	okay	@rosales_edder okay :)
3	There my baby just shared her dream	There my baby just shared her dream :)
4	so very true I do not count any of my follower...	so very true.\n\nI do not count any of my foll...
...	...	...
337	Hi This account is dedicated to kpop girl grou...	Hi! This account is dedicated to k-pop girl gr...
338	Good morning bestie! hope today brings you lo...	@Jefcfrey Good morning bestie😬\nI hope today b...
339	Kamek knows that single parent struggle Bowse...	RT @earthsong9405: Kamek knows that single par...
340	no way I love being mutuals with you youre wo...	@luvjoylovebot no way I love being mutuals wit...
341	new song Dominoes out June	RT @ansonseabra: new song “Dominoes” out June ...

Results

Translated Data Frame

	clean_tweet	spanish_tweet
0	the iraqi side was hella peaceful	el lado iraquí era muy pacífico
1	Make it happen then	Haz que suceda entonces
2	okay	okey
3	There my baby just shared her dream	Allí mi bebé acaba de compartir su sueño
4	so very true I do not count any of my follower...	Muy cierto, no cuento a ninguno de mis seguido...