Sequence to Sequence

(using Kor-eng dataset)

V2016120 김태형

Sequence to Sequence Model은 Text Generative Model중 대표적인 모델중에 하나입니다.

Sequence-to-sequence 학습 (Seq2Seq)은 하나의 도메인 (예 : 영어 문장)에서 다른 도메인의 시퀀스 (ex : 프랑스어로 번역 된 동일한 문장)로 시퀀스를 변환하는 모델 교육에 관한 것입니다.

```
"저것은 무엇인가요?" -> [Seq2Seq model] -> "What is that?"
```

Sequence to Sequence Model RNN계층이 2개가 사용되는데,
처음 RNN계층은 encoder 역할을 하여, 입력 시퀀스를 처리하고 자체 내부 상태를 반환합니다. 즉 다음 RNN계층 (Decoder)의 conditioning 역할이라고 볼수있다.
다른 RNN계층은 decoder 역할을 하여, 대상 시퀀스의 이전 문자가 주어지면 대상 시퀀스의 다음 문자를 예측하도록 훈련되게 만들만드는것이 핵심이며 인코더는 인코더의 상태 벡터를 초기 상태로 사용하며, 디코더가 생성하려고하는 내용에 대한 정보를 얻는 방법이다.

사용된 데이터는 아래주소에 가서 받았다.
http://www.manythings.org/anki/

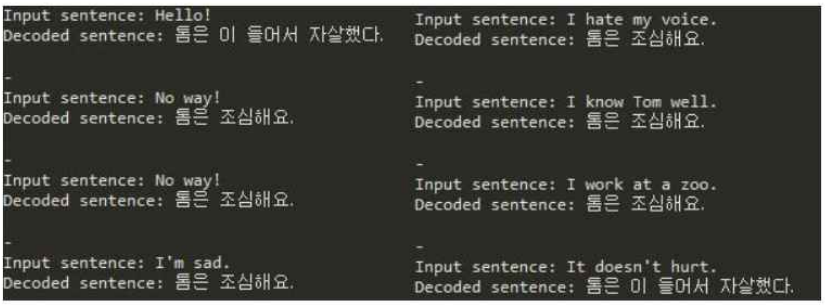
사용된 data는 Korean - English kor-eng.zip 를 다운받았으며,
알집을 풀고 확인해보면, 625개의 데이터로 구성되어 있으며 데이터형식은 아래와 같다.

```
I'll never tell you. 나는 결코 너에게 말하지 않을 것이다.
I'll take the wheel. 내가 운전할께.
I'm still in school. 저는 아직 학교에 다녀요.
I'm still in school. 나는 아직 학교에 다녀.
I've been kidnapped. 나 유괴 당한 적 있어.
It's not your fault. 네 탓은 아니야.
Let's do this later. 이건 좀 이따 하자.
Please drive slowly. 천천히 운전하세요.
That's your funeral. 그건 네 책임이야.
This is so relaxing. 이거 정말 기분 좋다.
This is so relaxing. 긴장이 풀리는데요.
Tom looks very sick. 톰은 많이 아파보여.
Tom must be winning. 톰이 이기고 있나봐.
Was the bank closed? 은행 문 닫혀 있었어?
Was the bank closed? 은행 문은 닫혀 있었어요?
Watch your language. 입을 조심해라.
Watch your language. 입조심해라.
What does this mean? 이게 무슨 뜻이에요?
Who taught you that? 그걸 누가 가르쳐 줬어요?
```

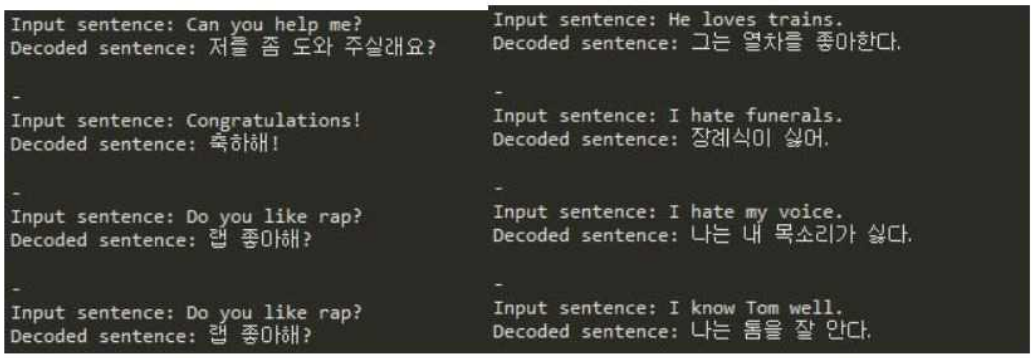
[kor-eng파일]

그리고 keras seq2seq 코드는
https://github.com/fchollet/keras/blob/master/examples/lstm_seq2seq.py 를 이용하였다.
기본적으로 파일을 넣고 돌려보면 error 'cp949' 오류가 뜰수 있는데 이 부분은 여기로 들어가서 확인후 고치면 될거같다.

위에코드에서 수정된 부분은 epoch 100 -> 1000으로 늘렸으며, latent_dim 256 ->1024 로 늘려서 확인해 봤다.
기본적으로 100번에 256으로 셋팅되어서 학습할 경우에는 학습이 아래와 같이 잘 안되는 부분을 확인할 수 있느넨,
epoch, latent_dim을 늘려서 확인해보면 괜찮은 결과를 얻을 수 있다.



epoch 100 / latent_dim 256
[결과]



epoch 1000 / latent_dim 1024
[결과]