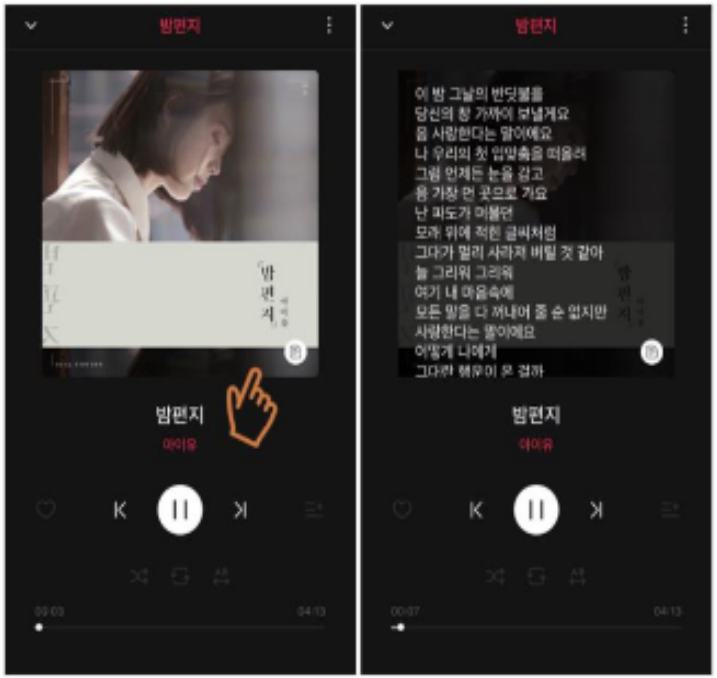
NLP 최종발표

박진우 이영신 유건희
조상현 유재형 하은겸

발표: 이영신, 박진우

Project 목표!!



GPT-2 model을 활용

문법에 맞고! 서사가 있는!

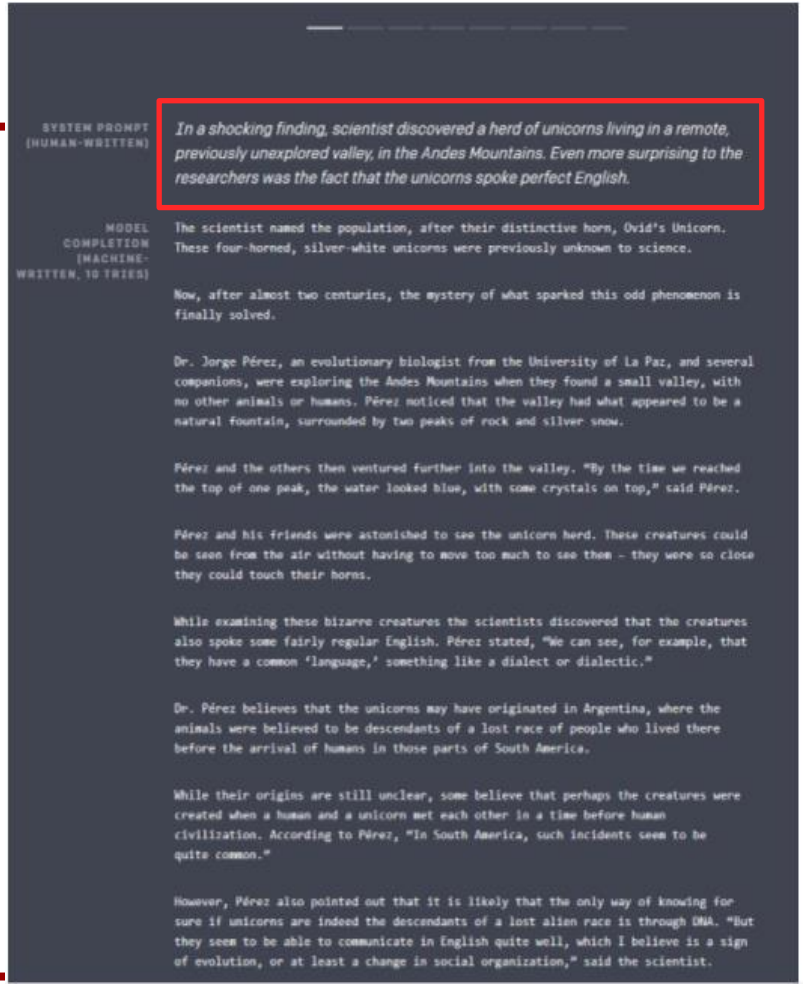
가사 만들기!!

1. GPT2
2. KoGPT2

GPT2

- GPT 모델의 다음 버전
- OpenAI에서 2019년 2월 14일 공개
- 악용될 소지를 고려해, Full model에 대한 소스 코드 공개 거부
  Light version만 공개

GPT2

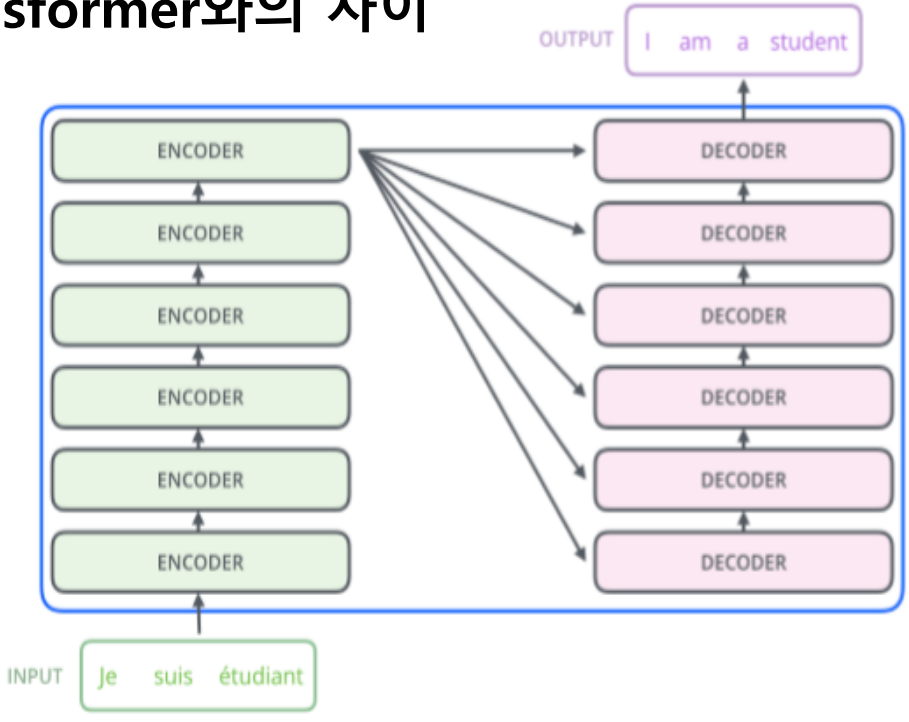

SYSTEM PROMPT (HUMAN-WRITTEN)
In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.
Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.
While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."
Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."
However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

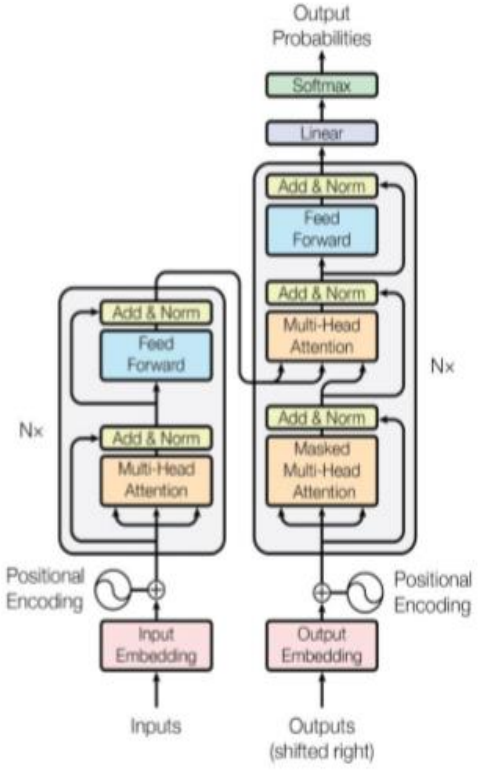
GPT2

- 새로운 architecture은 아님.
- GPT와 유사 decoder-only-transformer의 구조
- Bert나 GPT 보다 더 대용량의 데이터 셋을 이용해서 학습시킴. (40GB)

GPT2

Transformer와의 차이



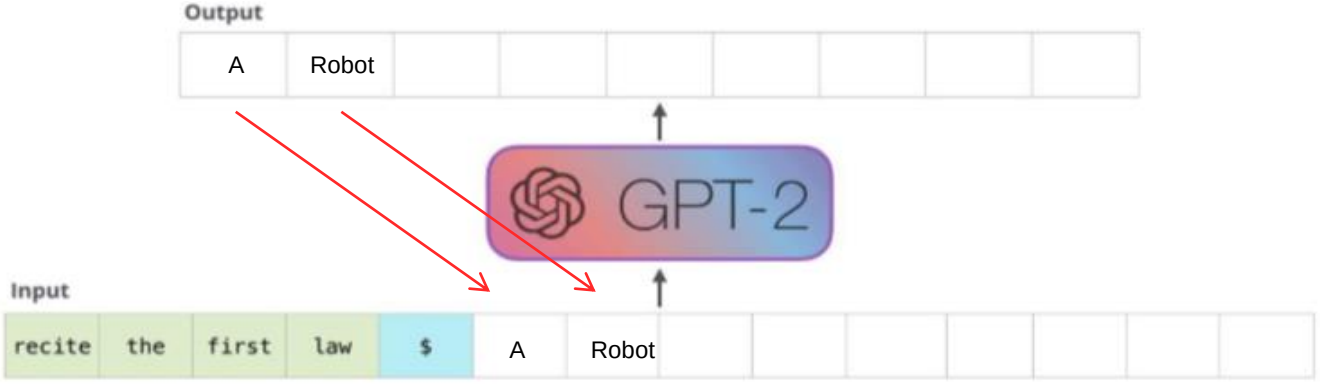


GPT2

BERT와의 차이

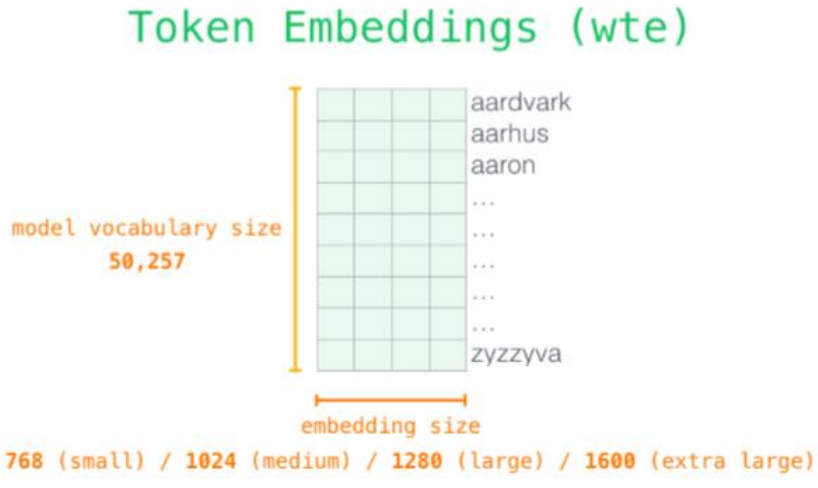
-GPT2는 auto regressive

-생성된 토큰은 그 다음 토큰을 생성하기 위한 input이 됨.

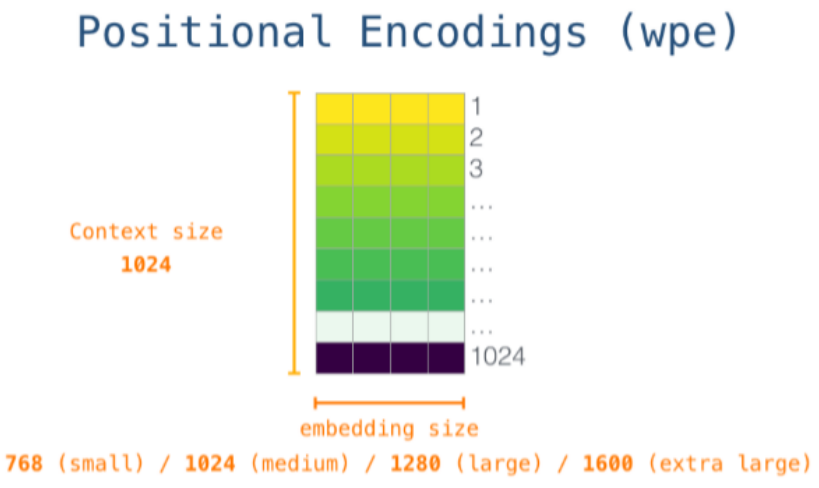


GPT2

Input encoding



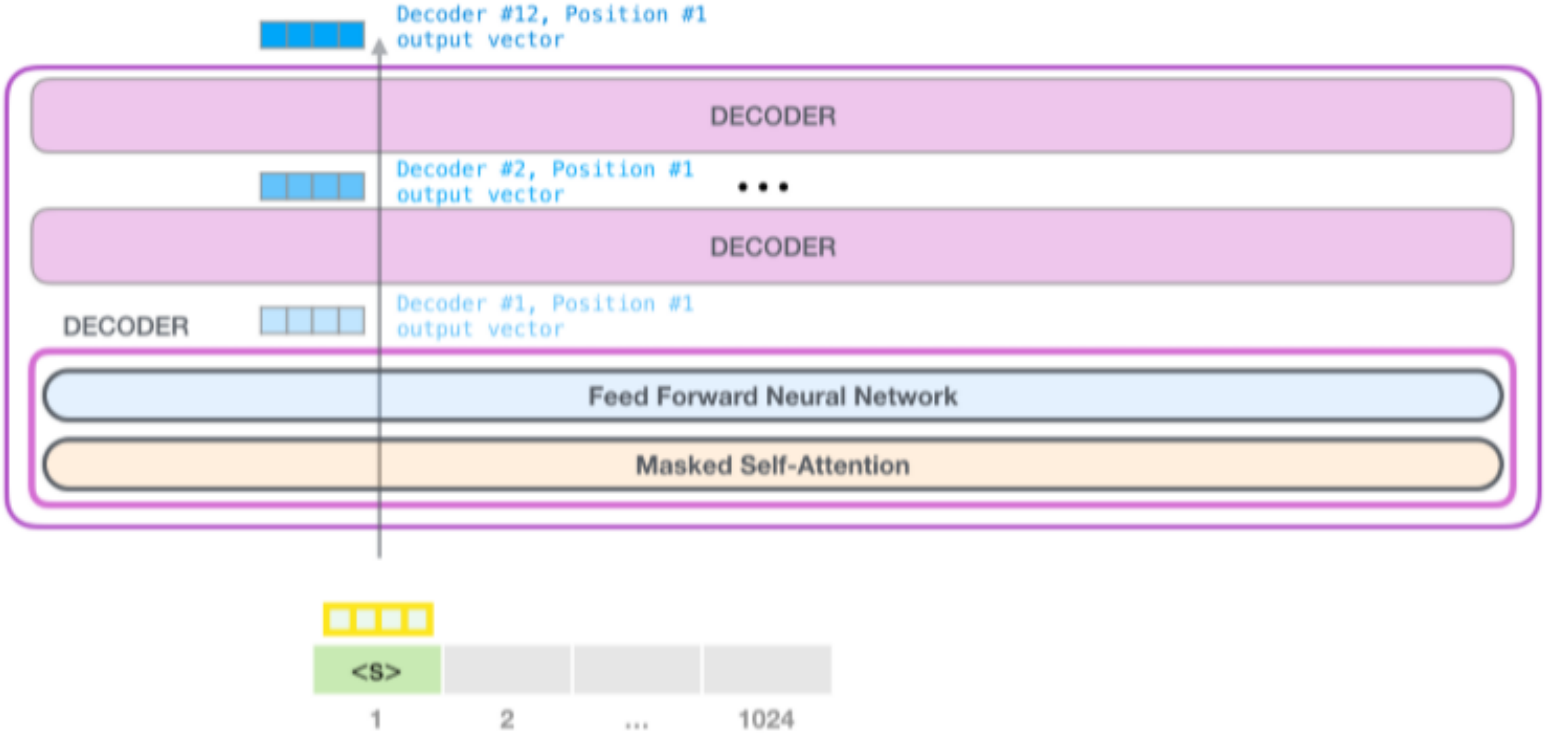
Positional encoding



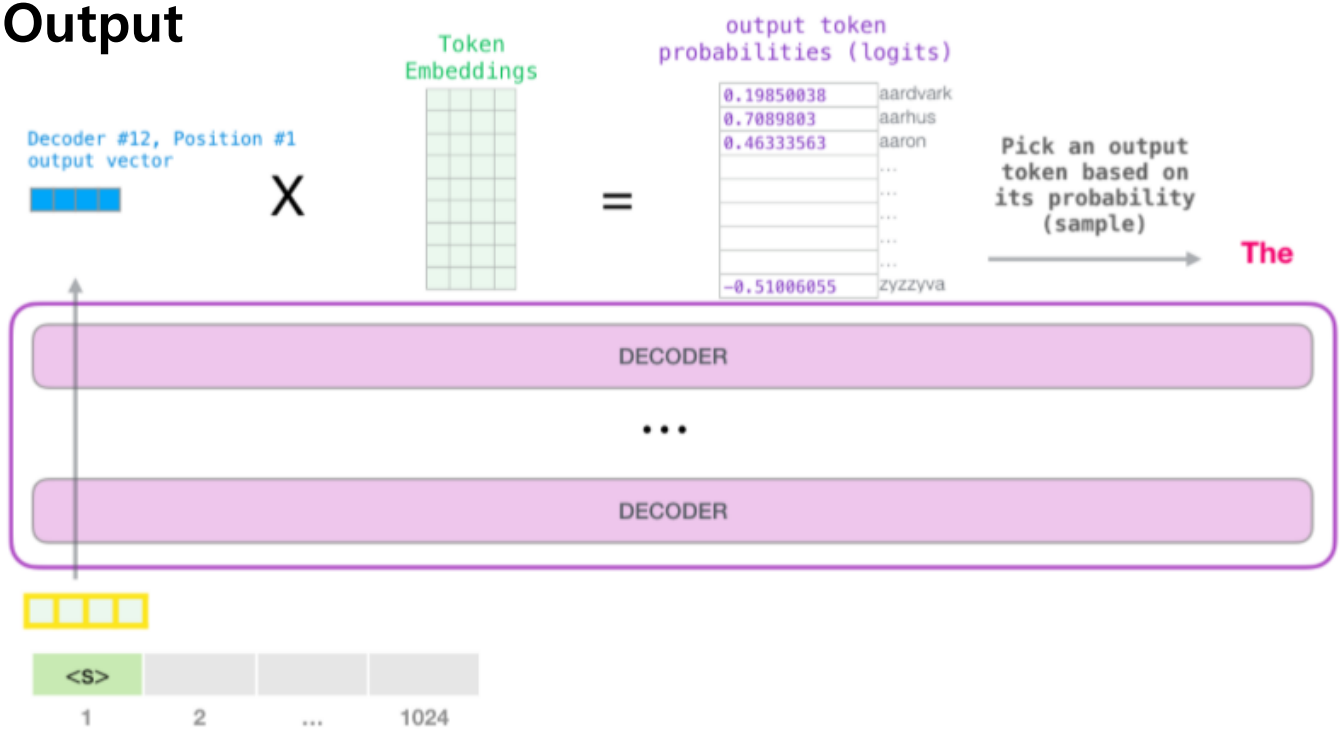
GPT2


DECODER
DECODER
=
Positional encoding for token #1
+
Token embedding of <s>
<s>
1
2
...
1024
Token Embeddings
Positional Encodings

GPT2


Decoder #12, Position #1
output vector
DECODER
Decoder #2, Position #1
output vector
DECODER
DECODER
Decoder #1, Position #1
output vector
Feed Forward Neural Network
Masked Self-Attention
<s>
1
2
...
1024

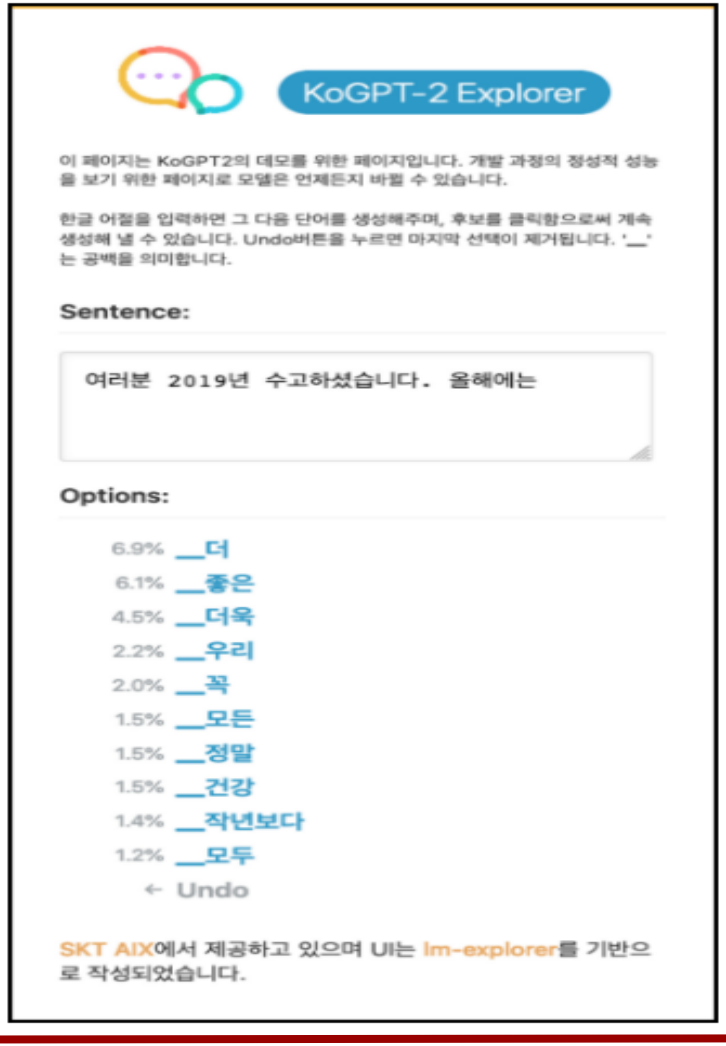
Model Output



KoGPT2

- 아마존웹서비스(AWS)와 SK텔레콤이 협력
- 한국어로 학습된 오픈소스 기반 모델
- OpenAI GPT-2 모델의 한국어 성능 한계 극복
- GPT2 base 모델
- 2천 5백만 이상의 문장으로 학습(wiki+news, 약 20GB)

KoGPT2


KoGPT-2 Explorer
이 페이지는 KoGPT2의 데모를 위한 페이지입니다. 개발 과정의 정성적 성능을 보기 위한 페이지로 모델은 언제든지 바뀔 수 있습니다.
한글 어절을 입력하면 그 다음 단어를 생성해주며, 후보를 클릭함으로써 계속 생성해 낼 수 있습니다. Undo버튼을 누르면 마지막 선택이 제거됩니다. '__'는 공백을 의미합니다.
Sentence:
여러분 2019년 수고하셨습니다. 올해에는
Options:
6.9% __더
6.1% __좋은
4.5% __더욱
2.2% __우리
2.0% __꼭
1.5% __모든
1.5% __정말
1.5% __건강
1.4% __작년보다
1.2% __모두
← Undo
SKT AIX에서 제공하고 있으며 UI는 lm-explorer를 기반으로 작성되었습니다.

3. Fine-tuning

Fine-tuning

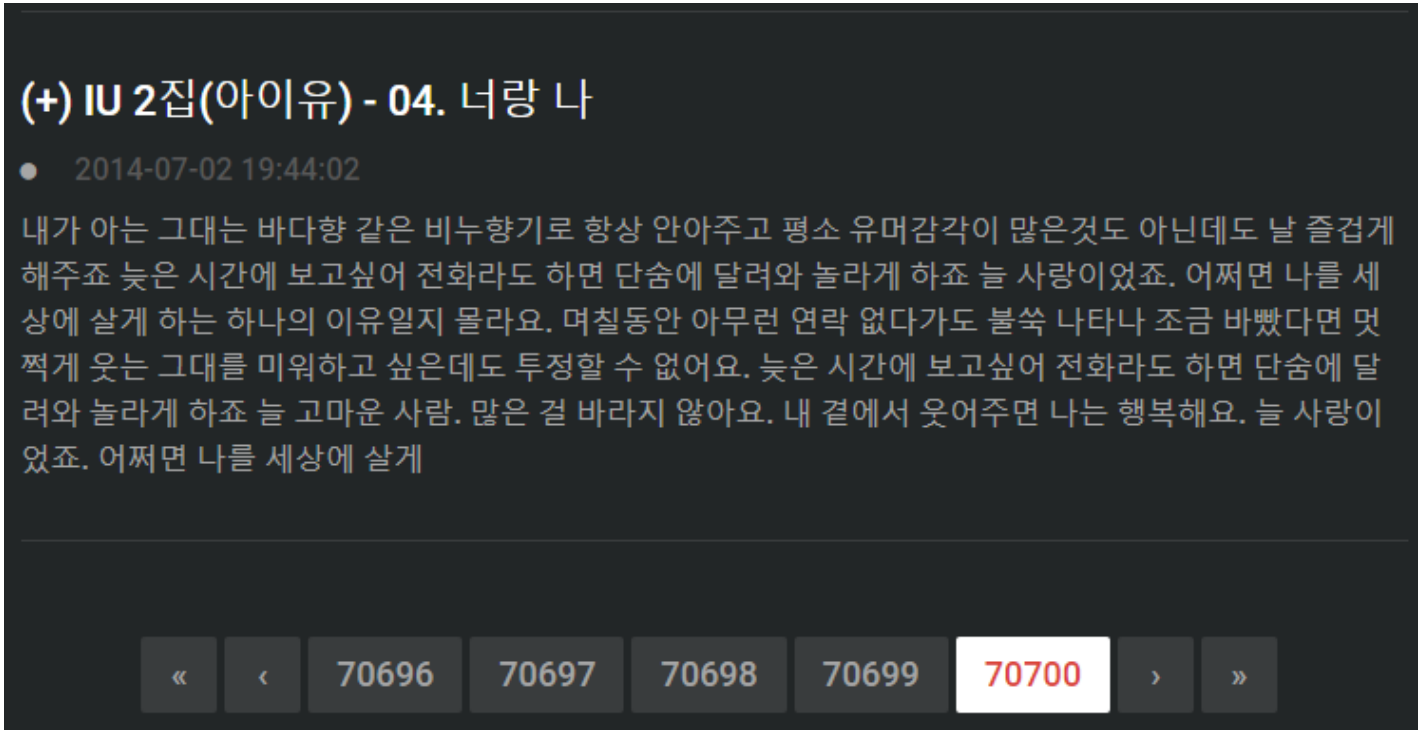
(2019년 한해를 보내며,)
새해에는 더 많은 사람들이
새해에 이루고자 하는
소망과 희망을 되새겨보는
시간이 **되었으면 좋겠다.'**

~다., ~한다 등
기사의 문체

문법에 맞고! 서사가 있는!

가사 만들기!!

Data


(+) IU 2집(아이유) - 04. 너랑 나
2014-07-02 19:44:02
내가 아는 그대는 바다향 같은 비누향기로 항상 안아주고 평소 유머감각이 많은것도 아닌데도 날 즐겁게 해주죠 늦은 시간에 보고싶어 전화라도 하면 단숨에 달려와 놀라게 하죠 늘 사랑이었죠. 어쩌면 나를 세상에 살게 하는 하나의 이유일지 몰라요. 며칠동안 아무런 연락 없다가도 불쑥 나타나 조금 바빴다면 멋쩍게 웃는 그대를 미워하고 싶은데도 투정할 수 없어요. 늦은 시간에 보고싶어 전화라도 하면 단숨에 달려와 놀라게 하죠 늘 고마운 사람. 많은 걸 바라지 않아요. 내 곁에서 웃어주면 나는 행복해요. 늘 사랑이었죠. 어쩌면 나를 세상에 살게
«
‹
70696
70697
70698
70699
70700
›
»

Data

남잔 왜 그래 사랑안해도
그런 말 참 쉽게 하네요
사랑해 그 말 뜻도 모르고
나의 맘을 여자 맘을 울리네요..
。ㄱr슴e 멈춘 ㅅr ㄹㅂ 。

┼☆ ＼。＼。SayCast [최강 감성파장 뮤。 직。 공 。간] ＼。＼。☆┼
┼☆ ＼。＼。＼。＼。＼。＼。RainY쁘띠에 ＼。＼。＼。＼。＼。＼。☆┼
┼☆＼。＼。＼。[비오는날의풍경II] 빗소리歌들린다 ＼。＼。＼。☆┼
┼☆ 최강뮤직 ☆┼
●┼┼Φ▶ ★º♡º 쁘띠에의 발칙한 n6n6 º♡º★ ◀Φ┼┼●
●┼┼Φ▶ ★º♡º 요조비은 ♡ BiOiOF7i º♡º★ ◀Φ┼┼●
2013.12.14 (사랑S러워서 사랑할 수 밖에 없는...~~^^*)

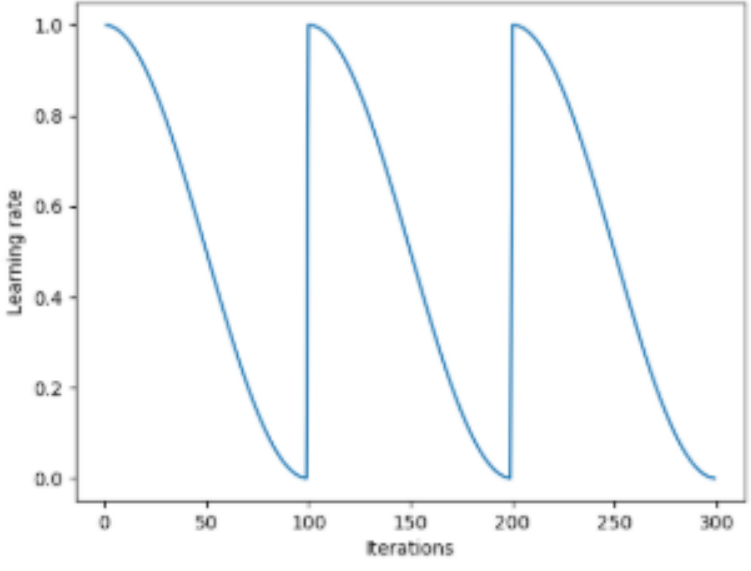
[첸] 一天一封?下的?心
yiti?n yif?ng xi? xia de zh?nx?n
하루 한 통 써내려간 진심
?藏在 最后一句的?意

한글,
영어

Fine-tuning

- ✓ Cosine Annealing

```
from tensorboardX import SummaryWriter
from utils import LineDataset
from torch.utils.data import DataLoader
from torch.optim.lr_scheduler import CosineAnnealingWarmRestarts
```



Fine-tuning

- ✓ AdamW optimizer

```
def arg_parse():
    parser = argparse.ArgumentParser()
    parser.add_argument('--learning_rate', default=0.000005, type=float)
    parser.add_argument('--optimizer', default='AdamW')
    parser.add_argument('--print_iter', default=20, type=int)
    parser.add_argument('--save_iter', default=500, type=int)
    parser.add_argument('--device', default='cuda')
    parser.add_argument('--gzip_loc', default='./excluded_dataset.pickle')
    parser.add_argument('--num_epoch', default=15, type=int)
    parser.add_argument('--batch_size', default=30, type=int)
    parser.add_argument('--check_point_path', default='./transformer_ckpt')
    parser.add_argument('--load_from_checkpoint', default=True, type=bool)
    parser.add_argument('--summary_url', default='./runs', type=str)
    parser.add_argument('--T_0', type=int, default=1)
    parser.add_argument('--T_mult', type=int, default=2)
    args = parser.parse_args()

    return args
```

4. Lyrics-

Lyrics-

(화려한 조명)속에 널 위한 side mirror girl
난 알고 있어 네 곁에서 너의 숨결이
마지막 날에 모든걸 다 잃어도 괜찮아
너를 위한 my love
listen to my love
그대여 내게 돌아와줘요
my love listen to my love
listen to my love
listen to my love
listen to my love

Lyrics-

운: 운전하면서
전: 전화를 하면
면: 면허가 정지되지
허: 허어어어~~~~~~~~~~~~~

(학점 잘 받고 싶어) 열정 하나로만 살아왔는데
일과를 다 뒤로 뒤로 미뤄 둔 채 학원으로 뛰어
대체 뭐가 어떻게 됐는지 좀 알 것도 같애
학원에 보낸 후 학원으로 뛰어
도대체 왜 이렇게 됐는지 왜 후회하는지 난 모르겠어
예이예 허우허우허우
예이예 허우허우허우허우
예이예 허우허우허우허우허우허우
예이예 허우허우허우
예이예 허우허우허우허우
예이예 허우허우허우허우허우

Lyrics-

(이상한거니) 이 번호로 전화해 봐
내 맘을 알겠니
이건 아냐 니가 누군데

무슨 말이 필요해 oh baby

이건 아냐 너를 위한 것
너의 사랑도
다 잊은 걸로 잊어볼래

너무도 멀리 멀리 가지마
다시 한번만 더 너에게로
니가 보여

너도 나 같은 건
다 잊은 걸로 잊어볼래

너무도 멀리 가지마
다시 한번만 더 너에게로
니가 보여

너도 나 같은 건
다 잊은 걸로 잊어볼래

너의 사랑도
다 잊은 걸로 잊어볼래

너무도 멀리 가지마
다시 한번만 더 너에게로
니가 보여

니가 너를 사랑한대도
나 혼자 남겨둘래

그 사람 사랑할래
너를 떠나볼래

니가 될래 널 보고 싶단 걸

Improvements

1. 전처리

2. Ovefitting문제

3. NLP Augmentation


야, 너두
작사 할 수 있어

마치며!