한국 드라마 특성을 가진 챗봇 제작

Making Chatbots with Characteristics of Korean Dramas

NLP팀

중앙대학교유승욱, 김상렬, 김중훈, 박경빈, 이하은, 임도연



팀원소개

About Our Team



유승욱 중앙대학교소프트웨어학부 데이터전처리,숏페이퍼및PPT작성



김상렬 중앙대학교컴퓨터공학부 KoGPT2챗봇파인튜닝



이하은 중앙대학교소프트웨어학부 데이터수집,숏페이퍼작성



김<mark>중</mark>훈 중앙대학교응용통계학부 Seq2seq챗봇파인튜닝



임도연 중앙대학교소프트웨어학부 데이터수집,발표준비



박경빈 중앙대학교소프트웨어학부 데이터수집,숏페이퍼작성

Part 1 | FAM MOG MISS

Topic Selection



지금은 한국 드라마 열풍!

Korean Drama



지금은 한국 드라마 열풍!

Korean Drama



Netflix와같은OTT서비스시장의확대로 한국드라마가더욱세계적으로뻗어나가고있음

드라마속 등장인물을 어디서나 느껴보자

Korean Drama



1. 사전 학습된 기존 챗봇 모델 이용

2. 파인 튜닝을 진행해 드라마 특성을 반영한 챗봇 제작

3. 글로벌로 뻗어나갈 한국 드라마 시장에 좋은 마케팅 요소가 될 수 있을 것

Part 2 | IMINITION 1

Data Definition and Collection



데이터 정의

Data Definition

세계인이 접할 수 있는 넷플릭스 드라마 中



- 1. 대중적으로인기를 끌어 많은 사람이 공감하였는가
- 2. 드라마특성상등장인물의 대화 형태가 뚜렷한가

데이터 정의

Data Definition



SKY캐슬(2018,20부작) 관전 포인트: 입시 경쟁 속에서 서로의 이기적인 모습



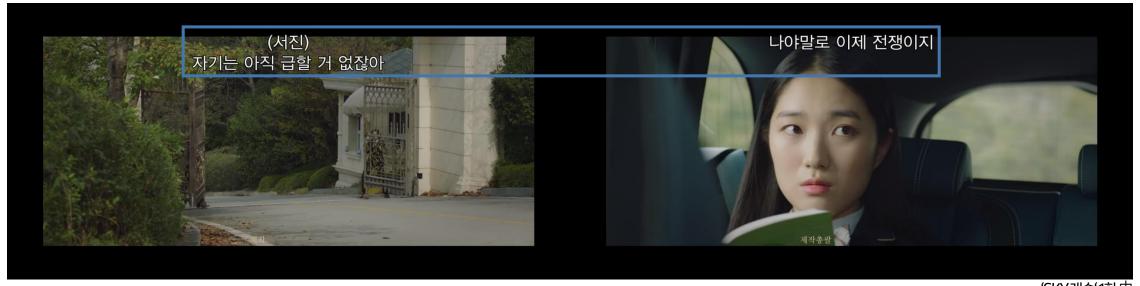
<u>미생</u>(2014,20부작) 관전 포인트: 직장인의 애환과 현대인 삶의 사실적 모습



<u>이태워 클라쓰</u>(2020,16부작) 관전 포인트: 불합리한 세상 속 청춘들의 저항적인 모습

데이터 수집

Data Collection

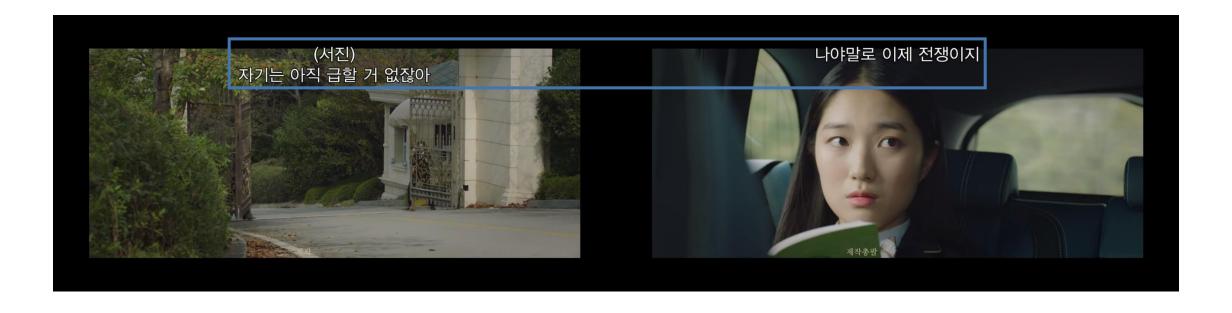


'SKY캐슬'1화中

- -한국어 대사에 해당하는 부분을 xlsx 파일로 추출함
- '(서진)'과 같은 대사를 말한 등장인물 정보, 시각적/청각적 정보를 나타내주는 정보, 비주요 등장인물들의 대사가 모두 포함되어 있어 추가 전처리가 필요한 상황

데이터 수집

Data Collection



Translation
(서진)자기는 아직 급할 거 없잖아
나야말로 이제 전쟁이지
걱정 마, 노하우 전수받으면자기한테 토스할게

xlsx파일로변환된모습

Data Preprocessing



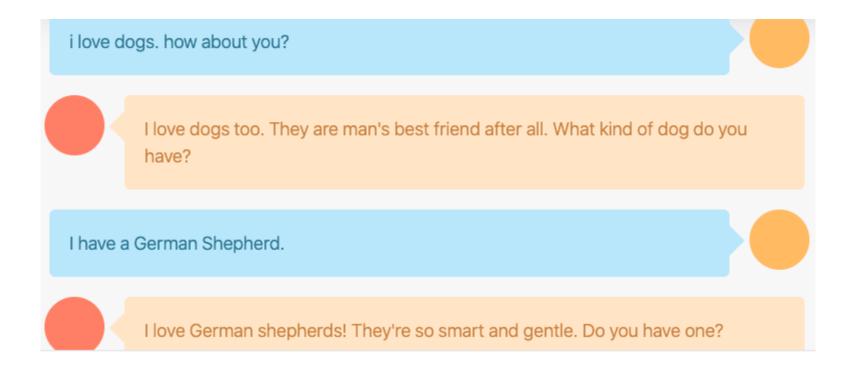
데이터 전처리

Data Preprocessing

- 기본텍스트전처리 한국어전처리 드라마전처리
- 제거 띄어쓰기검사 등장인물정보제거 노래 특수문자 제거 보래 특수문자 제거 자막검수자 정보 제거 대괄호 내용까지 통째로 제거

데이터 전처리

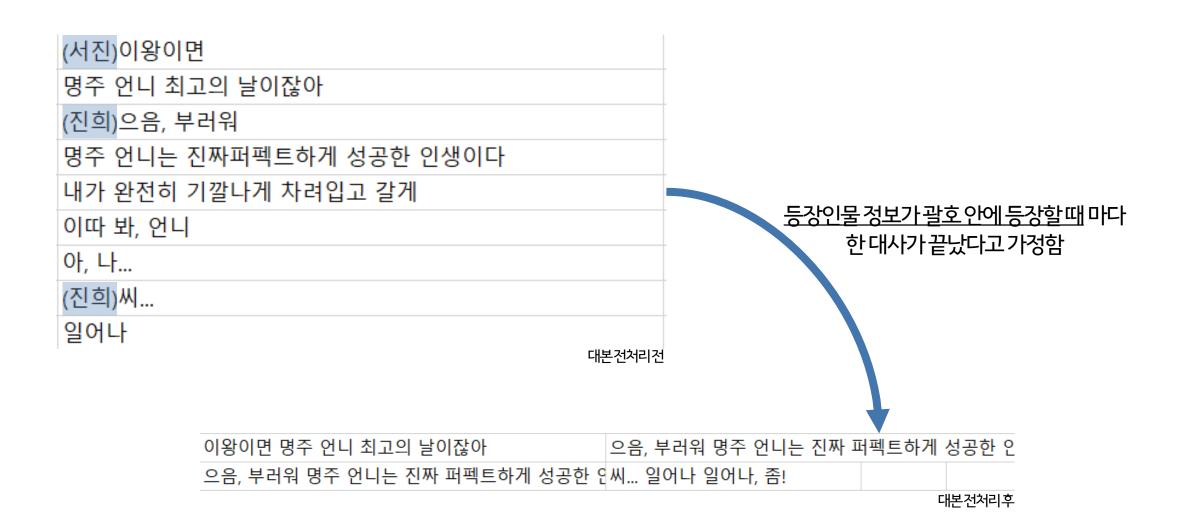
Data Preprocessing



챗봇 파인 튜닝에 활용하기 위해 드라마 대본 데이터를 '질의 형태'로 변경하였음

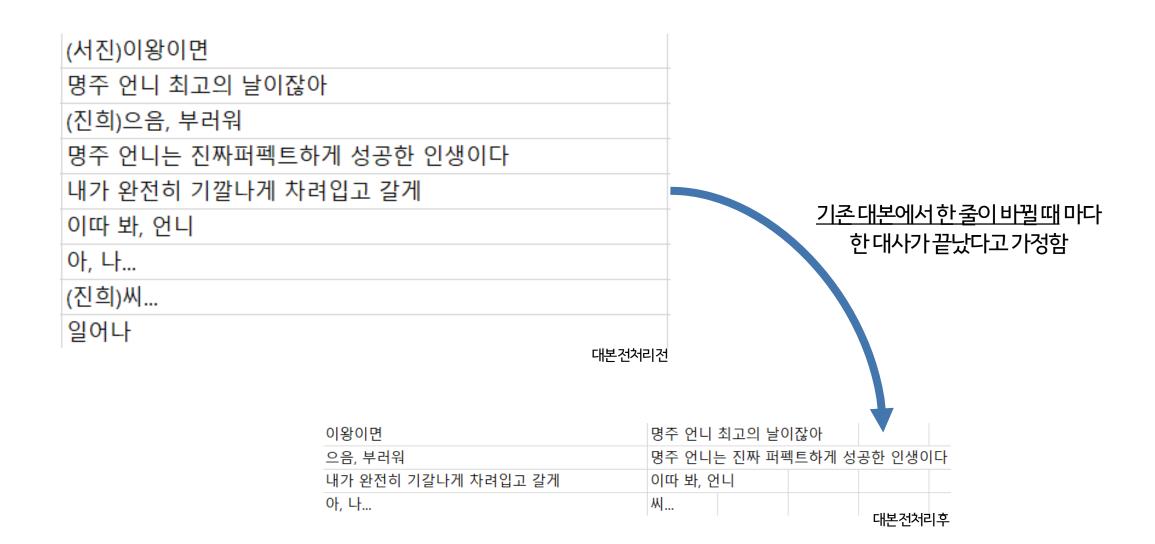
데이터 전처리: 질의 버전1

Data Preprocessing: QA Version 1



데이터 전처리: 질의 버전2

Data Preprocessing: QAVersion 2



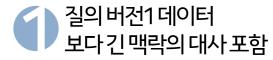
데이터 최종 형태

Generated Data Form

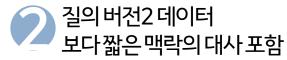
(서진)이왕이면
명주 언니 최고의 날이잖아
(진희)으음, 부러워
명주 언니는 진짜퍼펙트하게 성공한 인생이다
내가 완전히 기깔나게 차려입고 갈게
이따 봐, 언니
아, 나
(진희)씨
일어나

기존대본데이터

이왕이면 명주 언니 최고의 날이잖아	으음, 부러워 명주 언니는 진짜 퍼펙트하게 성공한 인
으음, 부러워 명주 언니는 진짜 퍼펙트하게 성공한 인	씨 일어나 일어나, 좀!



이왕이면	명주 언니 최고의 날이잖아
으음, 부러워	명주 언니는 진짜 퍼펙트하게 성공한 인생이다
내가 완전히 기갈나게 차려입고 갈게	이따 봐, 언니
아, 나	М



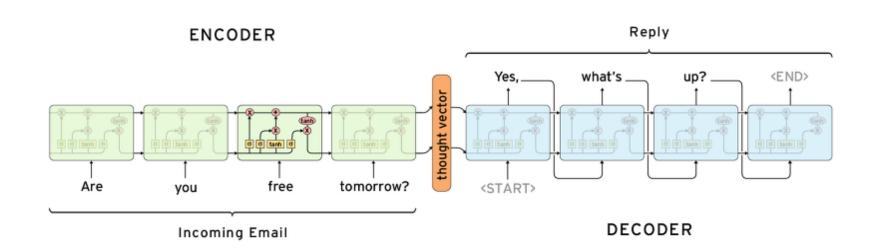
Part 4 | Seq2seq모델로챗봇파인튜닝

Chatbot Fine Tuning with Seq2seq model



Seq2seq 모델소개

About Seq2seq Model



- 인코더와디코더로이루어진 RNN 기반 모델로 기계 번역 및 챗봇에서 활발히 사용됨
- 출력 형태로 바로 이전의 입력까지만 고려하여 sentence generation의 정확도가 떨어질 수 있음
- <u>파인 튜닝된 챗봇 역시 문맥이 부자연스러운 모습을 주로 보여주었음</u>

(1) SKY 캐슬

'SKY Castle' with Seq2seq Chatbot



user > 중간고사 공부해야지
chatbot > 마유 아 이 제 안 되는 주제 에
user > 내가 너를 믿어도 될까
chatbot > 영재 엄마 가 합격 중 던져 주고 부모 영재 를 네 가 잘 되면 서준이 기준 이
user > 너 진짜로 서울 의대 가고 싶어?
chatbot > 김혜나 에 죽였다며 왜 죽였어 왜 도대체 왜 죽였어 여 네
user > 오늘 공부 다 했다
chatbot > 자 이 게
user > 나 쟤랑 같이 살기 싫어
chatbot > 야 너 엄마 어디 너 맞지 둘

(2) 미생

'Misaeng' with Seq2seq Chatbot



user > 우린 아직 다 미생이다 chatbot > 아니 이 거 못 모르는 사람 은 왜 안 들어 방금 회사 한테 받은 거 로 알 고 아니야 내 user > 네가 이루고 싶은 게 있다면 체력을 먼저 길러라 chatbot > 요즘 오늘 아주 맞게 그럼 장 백기 공장 에서 한 번 더 했어 왜 걔 가 암만 이 문서 에 서 user > 잊지 말자 나는 어머니의 자부심이다 chatbot > 이 거 지금 냐 고 user > 회사가 전쟁터라고? 밖은 지옥이다 chatbot > 과 사업 을 위해 서 얼마나 한국 의 상황 은 저번 에 처리 하고 싶은 겁니다 user > 모두가 그 길을 가질 수 있는 것은 아니다 chatbot > 그 래야 그만 마무리 해라 그 꼼꼼한 것 도 그만하면 병 이다 네 사장 님 저기 그리고 그

(3) 이태원 클라쓰

'Itaewon Class' with Seq2seq Chatbot



user > 내 가치를 니가 정하지마

chatbot > 검찰 출석 을 마치 고 돌아온 얼마 전 그날 저 는 장가 의 입 에서 직접 들었습니다 10년 원 에 대한

user > 니가 너인 것에 다른 사람을 납득시킬 필요 없어

chatbot > 저 도 제 같은 말 이야 할머니 이렇게 까지 있는 게 잘 못 된 일이 야 그렇게 살 고 백숙

user > 사람은 소신있게 살아야한다

chatbot > 왜 그 혜원 이 동생 이 서링

user > 사람이 있기 때문에 장사를 할 수 있는 겁니다

chatbot > 아이고 준 애 들 이 필요해 네

user > 해보고 판단해야지

chatbot > 이제 좀 죽였어 어디 가 제발 좀 써 주시겠어요 아 네 밥 도 씁 그냥 가시 죠 이제 부터

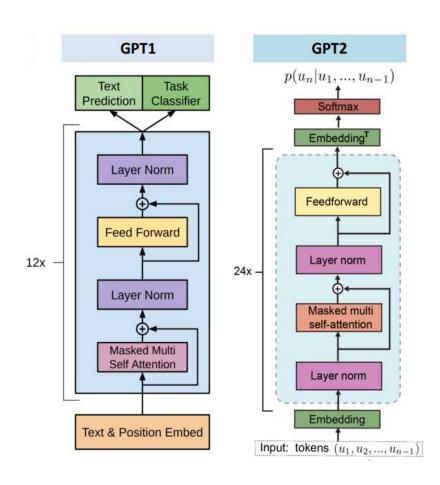
Part 5 | KoGPT2모델로챗봇파인튜닝

Chatbot Fine Tuning with KoGPT2 model



KoGPT2 모델 소개

About KoGPT2 Model



- GPT계열모델은머신러닝알고리즘을 활용해 구문론적, 문법적, 정보 등의 일관성을 갖춘 텍스트를 생성함
- 이전 단어들이 주어졌을 때 다음 단어가 무엇인지 추측하는 과정에서 학습되며, Transformer 구조를 응용한 언어 모델링 진행
- KoGPT2는 40GB이상의 대규모 한국어 말뭉치 텍스트로 학습됨
- <u>파인 튜닝된 챗봇에서는 Seq2seq 보다 자연스러웠지만,</u> <u>여전히 부자연스러운 부분이 존재하였음</u>

(1) SKY 캐슬

'SKY Castle' with KoGPT2 Chatbot



zeroshot	질의 버전1	질의 버전2
user>중간고사공부해야지		
bot>그래서너를사랑하지않아?	bot>예빈아우리딸잘먹고잘자고마 음편하게우리엄마인데	bot>그거말고그거
user>내가너를믿어도될까		
bot>무엇인가믿어요	bot>엄마저오늘은좀쉴게요기준이서 준이랑좀놀야줘야될거	bot>걱정돼서국는줄알았는데
user>너진짜로서울의대가고싶어?		
bot>뭐이런거는.	bot>난내실력으로만갈거니까예서 만떨어트리면돼요	bot>엄마가의대가라고해서의사됐 고싶지는않지만
user>오늘공부다했다.		
bot>공부다하면되니깐너를이해해!	bot>야하지마하지마공부안해도돼케 이야하지마공부안해도돼케이야	bot>엄마가하라니까더하기싫지?
user>나쟤랑같이살기싫어		
bot>저번에제가왔을때는너무기뻤지 만랑같이살기힘든게사실입니다!!!!!!!!	bot>엄마가편찮으시다니까아무래 도안되겠나봐요	bot>아빠도말씀하셨잖아

(1) SKY 캐슬

'SKY Castle' with KoGPT2 Chatbot

질의버전1데이터로실시간챗봇대화

```
(koepitadean) was depring a series of the BANDICAM -----
                                                                                                                                                                                                                 $ CLEAR VISITLE DIVICE
Distant
THEOLEGICAL TOTAL CONTRACTOR AND ADDRESS OF THE PROPERTY ADDRESS OF THE PROPER
ich iku eralen, aute volget grosefalen, autematik opstutiationehome, batch utvreän, beschmieten b
In characteristics from the result spot the stone, determined to Film, Atticked a particulations
annie. Liebt nat bitblioner is, tog every i stope EG, top you remore Home, is over Trop, to Fe-bi,
None, who emorphished a state of produce, noted parametrized the styrestic verticable -last chart.
dileador pod 'max cito pyris', no nedes t, no per oces t, no omity of dens, over it is to
repart that per pullering, and estation of the first training progress was a fresh fate Asse, reli
 ree grantles ship-Trace, i make been a selection to have a problem to wouldn't averbales, some baselment for
None, track grad memo-1, traditional a tronggraphy bett tenantage, val likely bukenyali tib, warman
HER support you't au'!
GEER I
```

(2) 미생

'Misaeng' with KoGPT2 Chatbot



zeroshot	질의 버전1	질의버전2	
user>우린아직다미생이다			
bot>오늘은서울잠실올림픽공원올림 픽홀에있는더플러스홀에서···	bot>아예이거랑또다른공장은?	bot>너는아직다	
user>네가이루고싶은게있다면체력을먼저길러라			
bot>"고당부했다.이에대해유 종 호교 수는 "평소처럼열심히···	bot>체력이약하면빨리편안함을찾 게되고그러면인내심이떨어지고···	bot>체력을 먼저 길러라	
user> 잊지 말자 나는 어머니의 자부심이다			
bot>라는생각을하였다. "여보! 너야. 우 리어머니께서…	bot>어머니는내게 자유를 주셨다 바둑 을 포기했을 때···	bot>자부심이다	
user>회사가전쟁터라고?밖은지옥이다			
bot>전세계적으로전쟁중인나라들 이너무나많으니까말이지.	bot>밖은지옥이다전무님	bot>전무님께서혹	
user>모두가그길을가질수있는것은아니다			
bot>그들은그길을택해내지못했을것 이다.	bot>야군대냐?회사에서무슨생명수 당이야	bot>이것때문이라는생각은안해본거 냐?	

(3)이태원 클라쓰

'Itaewon Class' with KoGPT2 Chatbot



zeroshot	질의 버전1	질의 버전2	
user>내가치를니가정하지마			
bot>그래야자랑이될수있고,더나아 가우리가사는모습을더잘볼수있다.	bot>어?정말로대표님을위하는마음 이조금이라도있다면그냥계속그렇 게가만이있고싶진않아요	bot>아	
user> 니가너인것에대해다른사람을납득시킬필요없어			
bot>또하나로는그냥나보고널싫어해?	bot>이새끼가이씨	bot>괜찮아	
user>사람은소신있게살아야한다			
bot>라는철학을갖고있다.그러나올 해노벨경제학상수상자는누가되든그는 	bot>결국넌퇴학난너나아저씨한테화 풀이를하고네표정이기대가된다	bot>소신있게살아야된다고가르쳐주 셨습니다	
user>사람이있기때문에장사를할수있는겁니다			
bot>라고대답한다.그가가게에들어서 자손님들은"아냐,우리손님들이손님들 이왔더라"며…	bot>그러니까지금저희한테투자한다 는회사도있으니까그돈을가지고와 서어?	bot>그래	
user>해보고판단해야지			
bot>이얘길하면서우리도이런분들 을많이만나고또이런분들많은사람들 이그부분에대해···	bot>그래근데너희가게경찰에신고 한사람그거나야재수없어	bot>그렇게시작하면돼	

Part 6 | 3E

Conclusion



결론

Conclusion

1. 드라마 대본에 드러난 일부 등장인물 정보의 부재

드라마 대본을 추출하는 과정에서 대사를 말한 등장인물의 모든 정보를 확인할 수 없어 임의로 질의 버전 데이터를 따로 구축한 부분이 모델 학습이 부정적인 영향을 미쳤을 것

<u>2. 드라마특성이 곧 챗봇으로 이어진다는 점에서 파인 튜닝의 의의 확인</u>

모든 상황에서 자연스럽게 학습된 것은 아니었지만, 기존의 사전 학습된 zeroshot 형태에 비해 파인 튜닝을 거친 챗봇은 드라마에서 언급된 표현들을 자유자재로 사용하였음

3. 차후 드라마 프로모션에서 중요한 역할이 될 가능성

한국 드라마가 세계에서 대중적인 인기를 끄는 만큼, 팬들에게 드라마 몰입을 위한 챗봇 서비스를 제공함으로써 드라마에 대한 기대감을 높이며 시청자층을 꾸준히 확보할 수 있음

Thank you!

Q&A

